Evolution of a Gene Regulatory Network Controlling Gut Patterning: Comparing Upstream Control and Chromatin Organization Around *Xlox/Pdx1* and *Cdx* genes in Sea Urchin, Sea Star and Amphioxus

Thesis

Danila Voronov, MSci

EVOLUTION OF A GENE REGULATORY NETWORK CONTROLLING GUT PATTERNING: COMPARING UPSTREAM CONTROL AND CHROMATIN ORGANIZATION AROUND XLOX/PDX1 AND CDX GENES IN SEA URCHIN, SEA STAR AND AMPHIOXUS

Doctor of Philosophy

School of Life, Health and Chemical Sciences

Affiliated Research Centre (ARC):

Stazione Zoologica Anton Dohrn, Italy

The Open University of London

December 2019

The work presented in this thesis was carried out in the lab of

Dr. Maria I. Arnone at the Stazione Zoologica Anton Dohrn, Naples, Italy

Director of Studies: Dr. Maria I. Arnone, Stazione Zoologica Anton Dohrn, Italy

External Supervisor: Prof. Peter W.H.Holland, University of Oxford, United Kingdom

ABSTRACT

ParaHox genes *Lox* and *Cdx* have a conserved role in embryonic development of many metazoan taxa. The two genes control the development of the gut and this control is done in time and location specific manner. The clustering of these genes on the chromosome is important for their temporal and spatial expression patterns. In the species with an intact cluster, the genes and their expression show temporal and spatial collinearity, while in the species where such cluster is broken, temporal linearity is altered. This raises questions as to the importance of genomic organization of these genes in the nucleus. This thesis employs four deuterostome species: *Strongylocentrotus purpuratus, Paracentrotus lividus, Patiria miniata* and *Branchiostoma lanceolatum*. In *S. purpuratus* and *P. lividus,* which are closely related sea urchin species, the clustering of ParaHox genes is absent. On the contrary, the sea star *P. miniata* and amphioxus *B. lanceolatum* genes *Lox* and *Cdx* are in a cluster along with *Gsx,* the third ParaHox gene, highlighting the importance of clustering for control of ParaHox genes. This thesis attempts to untangle the various factors controlling expression of the ParaHox genes and place them in the evolutionary context. The newly assembled genome for *S. purpuratus* has allowed to confirm that the cluster is, indeed, broken up in this species. HiC interaction data showed that the loci occupied by ParaHox genes in the sea urchin are not spatially close to each other even in the three dimensional organization of chromatin, unlike in the sea star or amphioxus that both have a tight cluster of ParaHox genes. In addition, chromatin accessibility assays, such as ATAC-seq, allowed to assess the open DNA regions and gain insight into their role in the expression of *Lox* and *Cdx,* which control development

of the sea urchin embryonic gut, suggesting that some of these open regions function as cis-regulatory modules (CRMs). Differential ATAC-seq and RNA-seq data at 48 and 66 hours post fertilization (hpf) for *S. purpuratus* revealed, which of the predicted cis-regulatory regions control gut development in the sea urchin, while single cell RNA-seq datasets for *S. purpuratus* 72 hpf pluteus allowed to filter out predicted transcription factors (TFs) and draft a gene regulatory network (GRN) for the three regions of the developing sea urchin hindgut at this time point. *In vivo* transgenic experiments, using reporter constructs, resulted in validating some of the predicted *S. purpuratus* CRMs and TFs, in particular showing a positive effect of SpHox11/13b on a *SpLox* CRM that overlaps *SpLox* transcription start site. This work also allowed to gain insight into the evolution of ParaHox gene control in closely related sea urchin species, through sequence comparisons and use of *S. purpuratus Lox* gene cis-regulatory modules in reporter constructs in *P. lividus*. The results of this analysis suggested CRM conservation and presence of the same transcription factor repertoire in homologous parts of the embryo in both species. Transcriptomic datasets, obtained for *S. purpuratus* embryos and tissues, highlighted the need for similar datasets from other species, in order to confidently untangle evolution of ParaHox gene control. Generated datasets allow for assessment of importance of chromatin organization for collinearity and set the foundation for future studies pertaining to evolution of the gut gene regulatory network in deuterostomes.

To my family

ACKNOWLEDGEMENTS

The current thesis represents the outcome of my three years of research at Stazione Zoologica Anton Dohrn during the PhD project. However, my journey to the PhD is much longer than that, so I would like to use this opportunity to express my gratitude to the people who were part of this journey.

First and foremost, I would like to say thank you to my mentor and supervisor Dr Maria Ina Arnone, for trusting me and allowing me to do this project, for the opportunities to learn and do cool and exciting stuff.

I thank Periklis Paganos, Jovana Ranđelović and Ines Fournon Berodia for the life in and outside the lab, which made me feel at home away from home, for jokes and pranks. I thank Dr Claudia Cuomo and Dr Elijah Lowe for guiding me in the lab, for training and teaching. Dr Giovanna Benvenuto for training and sharing the positive outlook on things. And, of course, for the work I performed with them.

I thank Elio Biffali, Pasquale De Luca and Elvira Mauriello for advice and technical wet-lab support. Marco Miralto for bioinformatics support and explanations of how such things work. Davide Caramiello for taking care of the precious animals.

I would like to say thank you to Dr Paola Oliveri, my Master's project mentor, for letting me enter the world of science. Dr Anna Czarkwiani for teaching and introducing me to molecular biology work. I thank Natalie Wood and other members of Oliveri lab for interesting discussions.

I am thankful to Prof Peter Holland for supervison and discussions which helped shape this thesis. To Dr Jose Luis Gómez-Skarmeta and Marta Magri for amazing collaboration on the ATAC-seq project. Dr Oleg Simakov for advice and teaching on bioinformatical analysis.

I am grateful to Charlotte Reynolds for being a great friend and for keeping me sane during tough times.

I thank Sergei Mendelevich Glagolev and the team of biology teachers at Moscow Highschool on the South-West N1543 for kindling my interest in biology and passion for research.

I thank my examiners for comments and suggestions that allowed to improve this thesis.

And last, but not least, I thank my family, whose presence and love I felt despite being far away in a different country.

I thank you all, without you this would have been impossible.

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

CHAPTER 1

INTRODUCTION

This chapter contains a general introduction to the thesis, introducing gene expression regulation, chromatin organization, gene regulatory networks and their components, as well as giving information on the experimental models used and their evolutionary relationships. This chapter also contains the aims and goals of this thesis.

1.1 Embryonic development depends on gene expression regulation

Any metazoan organism starts with a single cell - a fertilized zygote. Yet these single cells give rise to the vast diversity of metazoan body plans with multiple cell types, tissues and organs. The process of building these body plans from a single cell is embryonic development. In order to understand the diversity of the various cell and tissue types it is key to understand how they are built and how this process is controlled. Every cell in the organism has the same instruction set written in the DNA, with RNA molecules and proteins playing the role of effectors: building the cells and controlling cellular processes in the organism throughout development. Thus, cell and tissue type diversity depends on which genes are expressed within. Uncovering the mechanisms of gene expression control is essential to understanding of development and evolution. This control is facilitated by outside signals such as signalling molecules, internal regulators such as transcription factors or ncRNAs and the accessibility of RNA coding DNA regions to these internal regulators.

1.2 Chromatin organization controls gene expression

DNA in eukaryotic cells nuclei is packed into chromatin. Since DNA molecule is long, chromatin plays the function of condensing the genetic material to compact it inside nuclei. Naked DNA is around 2 nm across, however it is wrapped around histone proteins into nucleosomes. Around 147 basepairs (bp) of DNA is wrapped around core histones in each nucleosome, which are then connected together by linker DNA. This is a very dynamic beads-on-a-string structure that is 10 nm wide, it can be confined into topologically associated domains (TADs), which can

further condense by looping (Even-Faitelson et al. 2016). Both chromatin with its nucleosomes and other structures and proteins that bind DNA, such as RNA polymerases, are entities with physical dimensions and, in order for them to function, chromatin needs to provide access to DNA inside the nucleus. The 10 nm structure is very dynamic and represents open and active chromatin (euchromatin) that allows transcription machinery access, while the more condensed structures are closed and inactive (heterochromatin), as they prevent transcription machinery accessing the DNA, thus controlling gene expression.

In addition to DNA accessibility, three-dimensional organization of chromatin is crucial for bringing genomic elements and proteins together for transcription regulation. RNA polymerases bind to promoter regions of genes and drive transcription. Three-dimensional organization of chromatin has become even more intriguing in recent years, as it has been suggested that transcription occurs at transcription factories, complexes consisting of multiple proteins, such as transcription factors and polymerases (Dekker et al. 2013; Osborne et al. 2004). The location of these structures is believed to be relatively fixed in the nucleus, so chromatin looping and de-condensation allows to bring active genes to the transcription factory. Co-transcribed genes are brought together to the same transcription factory (Fraser et al. 2015). Functional arrangement of the chromatin in the nucleus is also concordant with an idea of gene segregation, which allows to separate active and inactive genes more efficiently, by grouping and separating them (Gaunt et al. 2003). Thus 3D organization of chromatin and relative position of genes is important for elucidation of mechanisms of gene expression regulation.

1.3 Epigenetic marks affect chromatin organization

Regulation of gene expression depends on protein-DNA binding between promoter regions and RNA-polymerase complexes. In addition to chromatin looping there are other factors that affect this binding. These include DNA modification, such as DNA methylation, and histone modifications.

DNA methylation, which is defined as methylation of cytosine, comprises around 5% of DNA cytosine content. Stretches of DNA with methylated cytosines are known as CpG islands. The majority of human genes have CpG islands upstream of them, methylation status of these islands is associated with proximal gene activity - methylated CpG islands are found near non-active genes (Vymetalkova et al. 2019; Talbert et al. 2019). Thus, methylated DNA is usually correlated with repression of gene expression.

Histone proteins play a crucial role in the chromatin organization of the nuclei of eukaryotic organisms. Some histone modifications open the chromatin and make it more accessible for transcription factors and transcriptional machinery such as RNA-polymerases. For instance histone acetylation H3K9ac by KAT2A is associated with open chromatin, same as histone methylation H3K27me3 or histone ubiquitination (Vymetalkova et al. 2019; Talbert et al. 2019). Other modifications such as methylation H3K9me3 are associated with closed and inactive chromatin (Vymetalkova et al. 2019).

All the above-mentioned factors contribute to chromatin accessibility to transcription machinery and reinforce the notion that chromatin organization is essential for gene control.

1.4 Cis-regulatory elements and transcription factors

Gene expression regulation can be viewed as either cis-regulation via cis-regulatory modules (CRMs) or trans-regulation via transcription factors (TFs). CRMs are regions of DNA that control genes located on the same chromosome as the CRMs themselves, such as enhancers or silencers. Cis-regulatory regions control the spatial and temporal gene expression patterns (Levine & Tjian 2003; Ong & Corces 2011) via transcription factor binding. CRMs are not very constrained in their location since they can be both upstream, downstream or inside the genes they control. In addition, they can be both proximal to the gene and distal.

Transcription factors are trans-regulating agents and are proteins that recognize a specific DNA sequence in promoters, enhancers or silencers and control gene expression. Transcription factors affect transcription by binding nucleosome modifying co-factors or by looping the chromatin, bringing enhancers closer to promoters and stabilizing RNA-polymerase complexes (Kostrewa et al. 2009; Ong & Corces 2011), allowing RNA-polymerases to bind DNA and transcribe RNA. Due to their crucial role in gene expression control, uncovering cis-regulatory elements around genes of interest and their associated transcription factors is key to understanding control of developmental processes.

1.5 Gene Regulatory Networks

The roles of cis-regulatory elements and transcription factors can be untangled in the context of gene regulatory networks (GRNs). In recent years use of GRNs

has become a common tool to facilitate study of regulation of cell and organism processes such as embryonic development, regeneration, cell differentiation and immune responses (Annunziata et al. 2014; Cholley et al. 2018; Roy & Kundu 2014; Singh et al. 2014). A gene regulatory network can be defined as a collection of genes that interact with each other to control cellular functions and processes. Thus reconstructing the gene regulatory network for embryonic development would shed light on the organization and control of this complex process, as interactions between the components of GRNs are instrumental in embryonic development (Annunziata & Arnone 2014).

Gene regulatory networks are represented by genes and their relative interactions between one another. Thus, a gene is a nodes of the network that is coding, for example, for a transcription factor and the effect of this protein on its target genes would be an example of interaction, such interactions are facilitated through CRMs, associated with the target genes.

A great effort to uncover the regulation of the *S. purpuratus* endomesoderm development was lead by Dr Eric Davidson (de-Leon & Davidson 2010; Tu et al. 2012; Peter & Davidson 2010; Davidson et al. 2002), who was also one of the initial contributors to the BioTapestry software used to draft GRNs (Longabaugh et al. 2005). Through this work many nodes of the sea urchin developmental GRN were uncovered, including *SpAlx1* (Ettensohn et al. 2003) and *SpEts* (Oliveri et al. 2008) in skeletogenic cells, *SpEse* (Materna et al. 2013) in the coelom, *SpBlimp1* (Livi & Davidson 2006) and *SpFoxA* (de-Leon & Davidson 2010) in endodermal cells, as well as their interactions with other nodes. The current

thesis aims to build the GRN of the digestive tract of the four deuterostome species, including *S. purpuratus*, and so expand on the GRN work of the many research teams that contributed to it. Thus part of this project is identification of putative CRMs and TFs that can recognize these CRMs and control transcription in tissues of interest, which, in context of this project, are the digestive system. This information would allow to draw the gene regulatory networks through revealing, which genes code for transcription factors, which genes are affected by the CRM, recognized by those transcription factors, and their interactions.

1.6 Chosen species and their evolutionary relationship

Evolutionary relationships between the phyla can also be uncovered by comparing the gene-to-gene interactions in the respective GRNs in each of these phyla. The gene nodes of these GRNs show that homologous genes may have different functions in different taxa, in addition, where the function of these homologous genes is the same among species, the wiring of the nodes within the GRN can be different taxa to taxa (Dylus et al. 2016). Studying the components and wiring of gene regulatory networks would give insight into the mechanisms of evolution of developmental programs. Evolutionary close species are especially suited for such studies since homologous genes are easier to find and comparisons are easier to make, which is the reason for choosing these close phyla for this project. The phyla of interest in this project are cephalochordates and echinoderms.

These two phyla are relatively close on the evolutionary tree (Lowe et al. 2015).The taxa in question are all deuterostomes, group defined by the anus

forming from the blastopore, while mouth forms in a secondary location. Other deuterostome characteristics traditionally include radial cleavage, indeterminate development and enterocoelic coelom. Echinoderms belong to the group *Ambulacraria,* which shares evolutionary roots with *Chordata,* a monophyletic group that includes cephalochordates, tunicates and vertebrates (Lowe et al. 2015). This project involved three echinoderms: *Strongylocentrotus purpuratus, Paracentrotus lividus,* both of which are euechinoid sea urchin species, and *Patiria miniata,* a sea star, along with a cephalochordate *Branchiostoma lanceolatum.* The project was mainly focused on the sea urchin species, in particular *S. purpuratus*, due to the availability of high quality genome assemblies, gene annotations and well established laboratory methods such as embryo culture (Adams et al. 2019), microinjections (Yaguchi 2019), transgenesis, including use of reporter constructs (Arnone et al. 2004; Nam et al. 2010), and gene perturbation techniques (Materna 2017).

Strongylocentrotus purpuratus, the purple sea urchin inhabiting the Californian coast of the Pacific Ocean, is an echinoid echinoderm with pentameral symmetry, globular body shape, multiple spines for protection and tube feet for mobility and feeding. First version of the *S. purpuratus* genome was published in 2006 (Sea Urchin Genome Sequencing Consortium et al. 2006) with updated versions published during the following years. This project makes use of the *S. purpuratus* genome version 3.1 (Cary et al. 2018; Sea Urchin Genome Sequencing Consortium et al. 2006), 816 megabases in size and the *S. purpuratus* genome version 5 (Unpublished, Dr Arnone lab is part of sequencing consortium for improved *S. purpuratus* genome), 844 megabases in size, which improves the

genome completeness and gives insight into chromosome level assembly of the genome. The *S. purpuratus* genome is thought to contain around 29072 coding genes many of which have homologs in vertebrates (Cameron et al. 2015; Cameron et al. 2009), with 21127 genes identified through transcriptome studies (Tu et al. 2012). These include sensory genes, biomineralization genes and genes, which have roles in human diseases, such as cancer and developmental syndromes (Sea Urchin Genome Sequencing Consortium et al. 2006; Kim et al. 2019; Jakubison et al. 2018; Yang et al. 2017).

Paracentrotus lividus is another echinoid species, local to the Mediterranean Sea, and, as such, the external characteristics of this species are almost identical to *Strongylocentrotus purpuratus,* except its colour, which is commonly brown, and spicule proportions. The genome is not yet published but was obtained from the authors (*P. lividus* sequencing consortium, Dr Arnone lab is part of sequencing consortium for *P. lividus* genome). The size of the genome is comparable to *S. purpuratus* assemblies as it is 880 megabases large and contains 30593 genes, 23573 have strong BLAST (Altschul et al. 1990) hits in *S. purpuratus,* so many genes are conserved between species (Malik et al. 2017).

Sea urchins have been used as organisms to study development for over a hundred years. Sea urchins are easy to keep in the lab and induce to shed gametes, they exhibit external fertilization and produce synchronously developing embryos, which are transparent, so formation of various cell and tissue types is convenient to observe. As with any metazoan, the sea urchin starts with a single fertilized egg, which then undergoes a series of cleavage divisions. Sea urchins

are deuterostomes, their cleavage is radial, holoblastic and reductive. The two first cleavage divisions are meridional and perpendicular to each other, the third cleavage is equatorial and, thus, also perpendicular to the first two. At the fourth cleavage stage animal pole cells divide equally and meridionally becoming mesomers, while the vegetal pole cells divide unequally and equatorially producing two sets of bigger and smaller cells. The bigger cells located closer to the animal pole are macromeres, while the vegetal-most cells are called micromeres. Later the mesomers divide giving rise to an_1 and an_2 derived cells, while the macromeres of the vegetal pole divide giving rise to veg_1 and veg_2 cells. The micromeres divide giving rise to four large micromeres and four small micromeres at this fifth division, while at the 6th division only the small micromeres divide. After the 6th division micromeres stop dividing until later stages. The cleavage divisions result in a morula, the cells of which keep dividing producing a blastula, which is a hollow ball of cells filled with fluid (blastocoel). At the blastula stage the cells start showing differences in anatomy, as the cells become ciliated on the outer sides of the blastula, the blastula becomes mobile and later hatches from the fertilization membrane, the vegetal pole cells become thicker forming the vegetal plate. The unequal fourth to sixth cleavage divisions are important for the sea urchin development since they determine the cell fates. The an_1 and an_2 cell populations will become the ectoderm, the veg_1 can give rise to both ectoderm and endoderm and the veg_2 will give rise to the endoderm and to the non-skeletogenic mesoderm, including coelom. The large micromeres will give rise to the skeletogenic mesoderm, while the small micromeres do not play a role in the sea urchin embryonic development but will contribute to

metamorphosis and to germ-line cells of the adult. The ectoderm is the outer epithelium of the embryo as well as most neurons, the endoderm is gut structures, while the non-skeletogenic mesoderm will form coelomic pouches, immune and muscle cells, with the skeletogenic mesoderm becoming the larval skeleton. As development continues the skeletogenic mesenchyme cells (also called primary mesenchyme cells (PMCs)) on the vegetal plate ingress into the blastocoel initiating the skeleton formation and leading the invagination of the other parts of vegetal plate into the blastocoel that results in the embryonic gut, the archenteron. This process is gastrulation. The archenteron elongates to the animal pole with the non-skeletogenic mesenchyme cells (also called secondary mesenchyme cells (SMCs)) appearing at its tip, until it reaches the region of the ectoderm that forms the mouth. All the mentioned structures continue to develop into the prism: the PMCs form skeleton, the SMCs give rise to the coelomic pouches, pigment cells, other immune cells and muscles, the gut becomes partitioned; which later develops into the pluteus (Figure 1.2). After feeding and dispersion the pluteus finds suitable location for metamorphosis and builds an adult body from its left coelomic pouch (Gilbert 2016).

Patiria miniata, the bat sea star, is an asteroid echinoderm with pentameral symmetry, star-like body shape, spines and tube feet. First assembly of the *P. miniata* has been published in 2012, and 30399 genes have been annotated in total (Cameron et al. 2015; Kudtarkar & Cameron 2017), many, again, showing similarity to the sea urchin species genes (Gildor et al. 2019). Its development is very similar to the one of sea urchin with the main difference that its cleavage division stays equal and micromeres are not formed, leading to a lack of

skeletogenic mesenchyme cells and absence of a larval skeleton. The sea star develops into a bipinnaria larva with a partitioned gut similar to the one in sea urchin prior to metamorphosis (Figure 1.2) (Flores & Livingston 2017; Gildor et al. 2017).

The annotations of the three echinoderm genomes and their similarity allows comparison between the functions and structures of genes involved in the molecular processes in these echinoderm species and gives evolutionary insight into the gene interactions in these taxons, their differences and similarities. The two sea urchins have diverged from each other 40 million years ago (Gildor & Ben-Tabou de-Leon 2015), while they diverged from the sea star around 581 million years ago (Figure 1.1). The echinoderm group has diverged from other ambulacraria 876 million years ago, and the whole ambulacraria diverged from chordates 896 million years ago (Figure 1.1) (Blair & Hedges 2005).

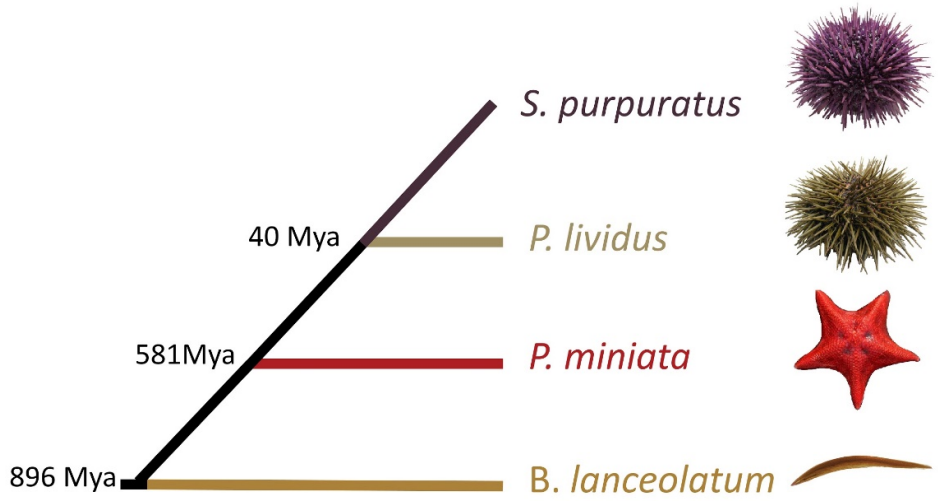


Figure 1.1 Evolutionary relationship between the chosen species

Branchiostoma lanceolatum is a cephalochordate, with a notochord, a nerve chord, segmented somites, a post-anal tail and a transparent fish-like body (Yu & Holland 2009).The *B. lanceolatum* genome, along with other sequencing data, became available in 2018 (Marlétaz et al. 2018). The genome is 475 megabases large with 21428 genes annotated. Cephalochordates are a sister group of vertebrates and tunicates which diverged from them 891 million years ago (Blair & Hedges 2005) and due to their unique position on the evolutionary tree, they are frequently considered to be basal chordates. Consequently, comparison between cephalochordate genomes and molecular mechanisms with vertebrate genomes and mechanisms allows to gain insight into the common ancestor of chordates (Yu & Holland 2009). The amphioxus also shows radial holoblastic cleavage, which starts with the first cleavage at the animal pole, the third cleavage is equatorial dividing the animal and vegetal poles. Cleavage leads to a morula formation. The morula then develops into a blastula. The tighter packed cells on the animal pole will become the ectoderm, while the looser larger vegetal cells will give rise to mesendoderm. The gastrulation starts with flattening of the blastula at the equator, this location will become the blastopore lip, then the mesoderm invaginates into the blastocoel making a layer under the ectoderm, and the blastopore becomes narrower marking the posterior end of the developing embryo. The dorsal and dorsolateral cells of the inner layer become the mesoderm, while the rest becomes the endoderm. The notochord starts forming at the gastrula stage by folding at the dorsal midline. The neural plate also starts forming then, later ectodermal cells move over the neural plate cells and fuse dorsally. This movement, along with neural plate curling, makes the

neural tube. During the elongation of the ciliated gastrula into neurula, the somites start to form, and, since amphioxus is a direct developer, further development of these structures and their rearrangement leads to the formation of the adult body (Holland & Yu 2004; Holland 2015).

The evolutionary proximity within echinoderms and between echinoderms and cephalochordate allows comparison of gene regulatory networks of these taxa to shed light on the evolution of deuterostomes. It is especially important, as this evolutionary position could be considered as a brink between chordates and non-chordates. Both echinoderms and cephalochordates have a well developed laboratory toolkits, which, along with the available molecular data and the relative ease with which these animals can be obtained and kept in the lab, make them good candidates for research into evolution of developmental processes. Combining molecular and sequencing data along with the information of the developmental processes enhances the evolutionary comparison (Cameron et al. 2015).

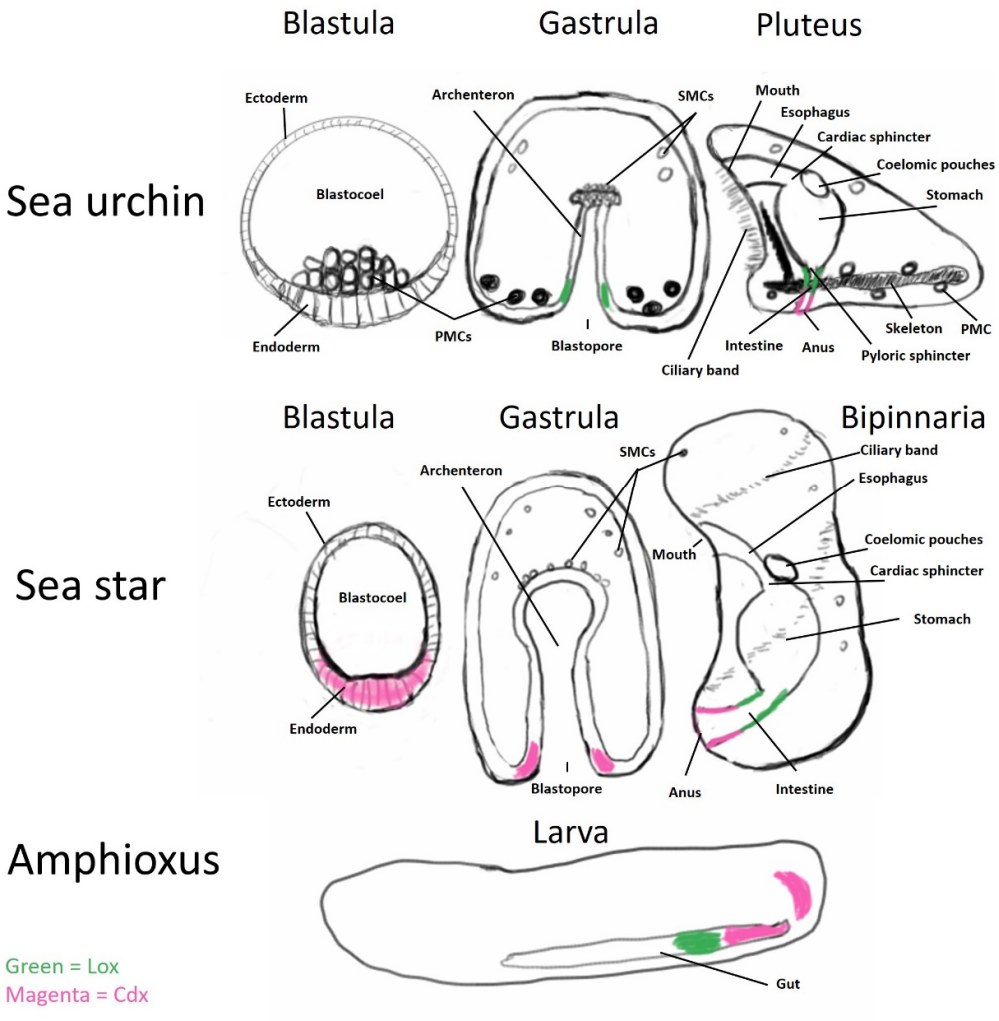


Figure 1.2 Schematic representation of embryonic stages of sea urchin, sea star and amphioxus. Spatial pattern of expression of ParaHox genes marked. Green is *Lox*, Magenta is *Cdx.*

1.7 Larval gut anatomy

Pentameral anatomy of the adult echinoderm body makes comparisons with bilaterally symmetric chordates difficult (Ji et al. 2012). However, most echinoderms have indirect development and go through a larval stage, which is capable of feeding and dispersion (Strathmann 1993; Annunziata et al. 2014). The larvae of both echinoids (plutei) and asteroids (bipinnaria larvae) show

bilateral symmetry and a tripartite gut, comparable to chordates (Annunziata et al. 2014). The echinoderm larval gut consists of an esophagus (foregut), a spherical stomach (midgut) and a tubular intestine (hindgut), separated from each other by sphincters (Figure 1.2). Cardiac sphincter separates the esophagus from the stomach, while pyloric sphincter separates the stomach from the intestine (Burke 1981) in the sea urchin. The esophagus serves to collect food particles into a bolus, which is then digested in the stomach, undigested food particles are passed on to the intestine and anus for excretion. These processes are facilitated by muscles in the esophagus and the sphincters as well as cilia throughout the digestive system (Burke 1981). Prior to these stages the gut at the gastrula stage is tubular without sub-sectioning, so the three partitions appear only at the prism stage as mentioned earlier.

Gut of the *B. lanceolatum* larva is also tubular then forming the adult gut with the embryo development (Figure 1.2). The digestive system of adult *B. lanceolatum* can also be broken into the three main parts. The foregut consists of mouth, located on the left side of the body, pharynx, consisting of pharyngeal slits and endostyle, which are used for filtering food from the surrounding water, and esophagus; the midgut comprises a stomach-like structure, which, however, does not exhibit bulging as in other deuterostome species, and hepatic cecum; the hindgut is separated from the midgut by a ciliated ilio-colon ring and consists of intestine and anus (Barrington 1937; Nakayama et al. 2019).

In addition to the anatomic similarity, the genes, instrumental to development of the gut in echinoderms and cephalochordates, are homologous to those in

vertebrates (Cole et al. 2009), such as *FoxA* (de-Leon & Davidson 2010), *Tgif* (Howard-Ashby et al. 2006), *Ptf1a* (Perillo et al. 2018), *Lox/Pdx1* or *Cdx* (Arnone et al. 2006), suggesting molecular similarity and evolutionary closeness of these gut development processes in echinoderms and chordates, which, along with similar anatomy, allows their comparison.

1.8 ParaHox genes control gut development

The two genes that play the pivotal role in the purple sea urchin gut development have been characterized: *SpLox* and *SpCdx.* Both genes are homeodomain transcription factors that belong to the ParaHox gene group along with *SpGsx* (Cole et al. 2009). ParaHox genes for the amphioxus have been characterized in 1998 and their putative role in the digestive system was suggested (Brooke et al. 1998). In bilaterians *Lox* and *Cdx* play a role in the central nervous system (CNS) development as well as the development of the endoderm (Perez-Villamil et al. 1999; Perillo et al. 2018; Metzis et al. 2018), including the digestive system. In the sea urchin and sea star development both of these genes control development of the hindgut from the pyloric sphincter through the intestine to the anus, although the sea star lacks a proper pyloric sphincter (Annunziata et al. 2019). Echinoderm *Lox* plays a role in the pyloric sphincter region development and the adjacent parts of intestine, while echinoderm *Cdx* plays a role in the development of the posterior-most region of the intestine and anus, role of echinoderm *Gsx,* on the other hand, is believed to be limited to neurons (Annunziata & Arnone 2014; Annunziata et al. 2014; Arnone et al. 2006). In the amphioxus larvae *AmphiXlox* (same as *AmphiLox*, sea urchin *SpLox* and

vertebrate *Pdx1* homolog) gene is expressed in the mid-posterior gut walls and transiently in two cells of the neural tube, *AmphiCdx* is expressed in the posterior-most region of the gut and this expression is continuous and extends into the posterior neural tube, while amphioxus *AmphiGsx* is only detected in the cerebral vesicle, region homologous to the vertebrate fore/midbrain (Brooke et al. 1998). All three genes have homologs in vertebrates, as there is one *Lox* homolog *Pdx1,* gene essential for pancreas development*,* three *Cdx* homolog genes *Cdx1*, *Cdx2* and *Cdx4,* involved in intestine development, and two *Gsx* homologs *Gsh1* and *Gsh2* (Coulier et al. 2000; Apweiler et al. 2004). In *S. purpuratus, P. lividus*, *P. miniata* and *B. lanceolatum* there is only one homolog of each ParaHox gene (Brooke et al. 1998; Arnone et al. 2006; Annunziata et al. 2013). Thus, it is worth noting that in *B. lanceolatum* ParaHox genes have not been duplicated, unlike in vertebrates, which is also supported by its genome and annotations. In general, vertebrate genomes are characterized by duplications, which make them distinctly different from invertebrate genomes (Holland 2003; Dehal & Boore 2005; Donoghue & Purnell 2005).

ParaHox genes form a group paralogous to Hox genes, and the two gene groups have arisen from the primitive ProtoHox cluster via duplication during the Cambrian explosion (Brooke et al. 1998). *Gsx* is paralogous to the anterior Hox genes, *Xlox* is paralogous to *Hox3* and *Cdx* is paralogous to the posterior Hox genes (Brooke et al. 1998; Arnone et al. 2006). The Hox genes code for homeodomain transcription factors that play a crucial role in the anterior-posterior patterning of the developing embryo through their linear expression patterns.

1.9 Genomic organization of Parahox genes and collinearity

An important feature of the Hox genes is their clustering on chromosomes in most bilaterians and, even though in some taxa this clustering is broken (Garstang & Ferrier 2013), the genes are located on the same chromosome adjacent to one another in a particular order (Gaunt 2015). In many organisms this synteny is correlated with the temporal and spatial expression patterns of the Hox genes, in other words, they exhibit temporal and spatial collinearity. Genes located at the 3' are expressed first and they are more anteriorly expressed, than genes that are located further to 5' that are expressed later and at more posterior regions of the embryo (McGinnis & Krumlauf 1992; Gaunt 2015; Holland 2013). In cases, where an intact cluster does not exist as in an acoel flatworm *Symsagittifera roscoffensis,* spacial collinearity is still observed, even though temporal collinearity depends on the Hox cluster being intact (Moreno et al. 2009).

In many taxa, such as vertebrates and amphioxus, ParaHox genes are also arranged within clusters, which also confer spatial and temporal collinearity of expression for these genes. In the amphioxus species *Branchiostoma floridae* ParaHox genes are located on one chromosome with *AmphiGsx* being 5'-most gene, followed by *AmphiXlox*, in the same orientation, and then by *AmphiCdx*, in the opposite orientation. *AmphiGsx* and *AmphiXlox* are separated by 25 kilobases, while *AmphiXlox* and *AmphiCdx* are separated only by 7.5 kilobases (Brooke et al. 1998). In chordates *Cdx* is expressed first at mid-gastrula stage around the blastopore, expression of *Xlox* starts later in the posterior endoderm, then localizing to the midgut and neural tube, with *Gsx* being the gene, whose

expression starts last and only in the neural tube. Spatially, *Gsx* is the anterior gene, *Lox* is the "middle" gene and *Cdx* is the posterior gene.

Similar situation has been shown in an echinoderm, bat sea star *P. miniata,* where the ParaHox genes are also clustered and show expression patterns similar to chordates (Annunziata et al. 2013) both in location and timing. *PmCdx* is the 3'-most gene and it is the most posterior in its domain of expression, while being the first of the three ParaHox genes to be transcribed with expression starting at 20 hours post fertilization (hpf) peaking at 24 hpf around the blastopore and at 72 hpf at the hindgut and anus, keeping this posterior location further into embryonic development. *PmLox* is the middle gene both on the chromosome and in the embryo, with its spatial and temporal expression detectable at 48 hpf ectodermally while its endodermal expression starts at 52 hpf at the posterior end of the archenteron; at later stages *PmLox* expression is confined to midgut-hindgut boundary. *PmGsx* is located at the 5' on the chromosome, however its expression was not detectable during early development after the egg stage, in which maternal mRNA of *PmGsx* is present, although *PmGsx* may have potential role in late development and metamorphosis into adult body (Annunziata et al. 2013). On the chromosome *PmGsx* and *PmLox* are only 31 kilobases away from each other, while *PmCdx* is 13 kilobases away from *PmLox*. Orientation of ParaHox genes in *P. miniata* follows ParaHox genes orientation in amphioxus, with *PmGsx* and *PmLox* being in the same orientation, and *PmCdx* in opposite (Annunziata et al. 2013).

However, sea urchins, such as *S. purpuratus,* do not have an intact cluster since in *S. purpuratus* genome version 3.1 the three genes are located on different scaffolds with no information on whether these scaffolds are adjacent to each other. The sizes of these scaffolds suggest that there is no ParaHox cluster in *S. purpuratus* (Sea Urchin Genome Sequencing Consortium et al. 2006; Kudtarkar & Cameron 2017). Lack of an intact cluster in *S. purpuratus* is associated with the loss of temporal collinearity, as it appears reversed, as *SpGsx* is the first of ParaHox genes to start expression in *S. purpuratus.* Expression of *SpGsx* starts at around 24 hours post fertilization, as its transcripts can already be detected at this stage, followed by *SpLox* at 32 hours and then by *SpCdx* at 40 hours post fertilization. Spatially *SpGsx* is confined to ectodermal cells, which is consistent with the role of this gene in development of the nervous system, which is derived from the ectoderm (Arnone et al. 2006). Both *SpLox* and *SpCdx* are mostly endodermal, as they are expressed in the archenteron from the gastrula stages. At later stages, such as pluteus from 72 hours post fertilization *SpLox* is found to be expressed in the intestine and pyloric sphincter region as well as neuronal cell populations, *SpCdx* at these stages is also found in the hindgut but closer to the anus, which is posterior to the *SpLox* expression regions. The expression locations of the two gut associated genes overlap however, as *SpCdx* expression can be described as somewhat of a gradual increase in expression to the posterior-most end of the embryo (Arnone et al. 2006). Therefore, spatially, *SpGsx* can still be considered an anterior gene, *SpCdx* the posterior gene and *SpLox* to be the gene expressed at the “middle” of the embryo. Thus, dispersed genes of the ParaHox group still show a spatial collinearity, similar to chordates

and asteroid echinoderms, and a reversed temporal linearity, compared again to chordates and asteroids (Annunziata et al. 2013; Arnone et al. 2006). This suggests that clustering is important for temporal aspects of gene expression, but may be less important for spatial expression patterns. This notion is also supported by data from other species such as *Ptychodera flava,* an acorn worm, which is evolutionary close to echinoderms (Ikuta et al. 2013). It has an intact ParaHox cluster, with the genes in the same orientation as in the chordate ParaHox cluster, however, unlike chordates, *P. flava* exhibits temporal collinearity of expression of its ParaHox genes, but no spatial collinearity (Ikuta et al. 2013).

There have been a number of ideas proposed to the significance of collinearity (Gaunt 2015). One of such ideas is enhancer sharing. As mentioned earlier enhancers are regions of DNA that increase gene transcription by binding proteins such as transcription factors. Enhancers are cis-regulatory elements that are in close proximity to the genes they regulate. Three-dimensional location of enhancers is crucial for bringing them near the target gene, as these DNA elements may be thousands of base-pairs away from the gene (Annunziata et al. 2013; Shlyueva et al. 2014). Confining all ParaHox genes into a cluster would allow these genes to share enhancers, which can facilitate their activation in the observed order, as it would bring the genes physically close on the chromosome. Enhancer sharing is also possible for distant genes provided they interact in 3D conformation of the chromatin. To gain insight into potential enhancer sharing between distal genes a techniques like 4C or HiC can be employed (Ulianov et al. 2015; Dekker et al. 2013), highlighting importance of chromatin organization for expression control in the context of synteny and collinearity.

1.10 Known components of gut GRN

Transcription factors that bind to regulatory regions of DNA control gene expression. Interaction between these factors and their targets is a gene regulatory network (Lowe et al. 2016). A substantial part of the GRN controlling gut development in *S. purpuratus* and a part of *P. miniata* GRN have been reconstructed previously (Cole et al. 2009; Annunziata & Arnone 2014).

In sea urchin two genes have been identified as part of GRN controlling foregut in sea urchin: which are *SpBrn1/2/4* and *SpFoxA* with *SpFoxA* being a node in all regions of the gut: foregut, midgut and hindgut.

The stomach GRN nodes throughout development consist of transcription factors *SpBlimp1*, *SpFoxA*, in addition to mannose receptor *SpManrC1A* and calcium binding *SpCabpf (SpEndo16),* which are frequently considered stomach terminal differentiation genes.

The pyloric sphincter at 66 hpf and 72hpf GRN contains *SpBlimp1*, *SpLox*, *SpCabpf* and myosin heavy chain *SpMy18A* (*SpMhc)* at later stages when the pyloric sphincter is fully formed. *SpBlimp1* is shown to be activator of *SpLox*, which in turn activates *SpMhc* and *SpManrC1A* in the stomach through an unknown intermediate.

The intestine GRN from 66 hpf to 72 hpf contains of *SpBlimp1a*, *SpLox*, SpCdx and *SpFoxA* with SpBlimp1a activating *SpLox*, SpLox activating *SpCdx*, while SpCdx in turn inactivates *SpLox,* restricting it to the pyloric sphincter and the anterior parts of the hindgut, and *SpCabpf. SpCdx* also has a positive

autoregulatory loop. In the anus *SpBlimp1a*, *SpHox11*/13b, *SpBra*, *SpCdx* and *SpWnt10* make up the nodes of the anal GRN at the same time points. SpLox, SpHox11/13b, SpFoxA and SpBra are known to have a positive effect on *SpCdx* expression, which in turn has a positive regulatory effect on itself, *SpBra*, *SpWnt10* and a negative effect on *SpCabpf* which it also does not colocalize with. SpHox11/13b also plays a role of activator of *SpBra*, early on in development, which suggests its role as an early activator of the endodermal lineage leading to gut formation, while SpWnt10 shows negative control of *SpLox* expression in posterior-most gut regions (Cole et al. 2009; Annunziata & Arnone 2014; Annunziata et al. 2014).

Prior to the development the gut tissue at the blastula stage genes *SpOtx* and *SpGatae* play roles in the activation of *SpBra*, *SpGatae* and *SpBlimp1* as well as other gut genes of the GRN, while *SpBlimp1* also has a positive effect on the expression of these too genes as well. *SpOtx* has an autoregulatory loop and also activates *SpFoxA*. Other genes involved in the specification of the gut tissues at different stages of sea urchin development include *SpTgif*, *SpHh*, *SpDac, SpKrl*, *SpEve* and *SpMyc*. Genomic view of the known gut GRN up to 72 hours is presented in figure 1.3 adapted from http://www.echinobase.org/endomes/ and Annunziata and Arnone 2014.

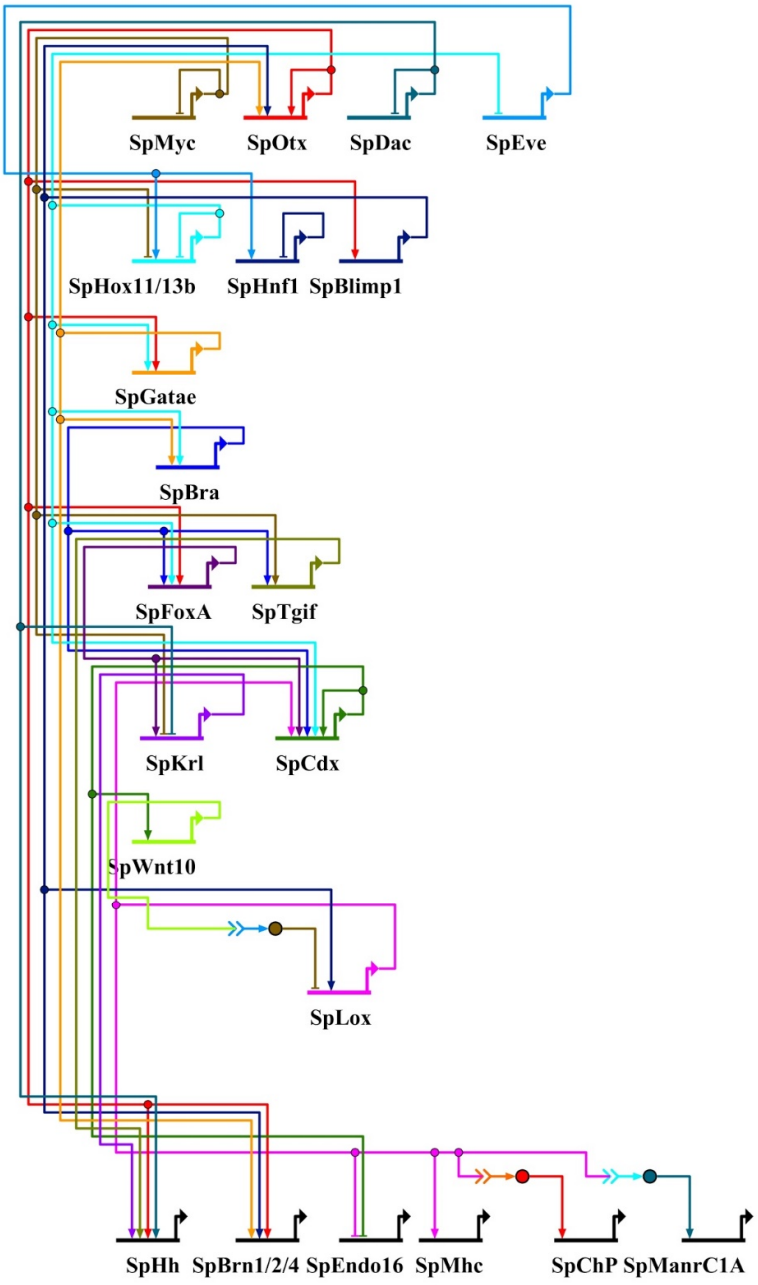


Figure 1.3 Known components of *S. purpuratus* gut GRN up to 72 hours post fertilization

Except a few known direct interactions such as SpBra driving *SpFoxA* expression and SpGatae, SpBlimp1 and SpOtx driving *SpOtx*, for majority of interactions stated in the known gut GRN it is unclear whether they are direct or not, with the interactions being far from numerous, which highlights the need to uncover the regulation of ParaHox genes by elucidating, which transcription factors have direct effect on them, as well as identifying their associated CRMs. Many transcription factors are shared between gut development of *S. purpuratus* and *P. miniata* (Lowe et al. 2016). This known GRN is a starting point for further analysis to determine the possible function of the forementioned transcription factors and the evolution of the GRN (Lowe et al. 2016; Lowe et al. 2017; Lowe et al. 2019) upstream of ParaHox genes.

1.11 Aim of the thesis

The main aim of this thesis is to elucidate the mechanisms that control the expression of *Lox* and *Cdx* genes in the four species stated: *S. purpuratus*, *P. lividus*, *P. miniata* and *B. lanceolatum*. As stated above the main actors controlling gene expression are chromatin three-dimensional organization, chromatin accessibility to transcription machinery, transcription factors and their associated CRMs. Therefore the goals of this thesis are to deduce genomic organization of ParaHox genes in the four species: the relative locations of ParaHox genes on the chromosomes and the three-dimensional organization of the loci that ParaHox genes *Lox* and *Cdx* occupy, to assess accessibility of the genomic DNA near these loci to identify putative CRMs and to identify transcription factors that act on these CRMs thus driving expression of ParaHox

genes in the spatial and temporal manner observed. This information allows to reconstruct the GRNs upstream of ParaHox genes. In addition, studying these four actors in the species of interest would not only shed light on the development of digestive systems in these taxa but also give evolutionary insight on developmental systems and gene control. These are the aims and goals of this thesis, which will be further addressed in the consecutive chapters.

CHAPTER 2

MATERIALS AND METHODS

This section contains materials and methods that were used for the project. Methods range from spawning animals and rearing larval cultures to preparing sequencing libraries and analyzing sequencing data. Methods not described elsewhere are described here in detail.

2.1 *In vitro* fertilization and embryo culture

Echinoderms do not display sexual dimorphism, therefore their sex can only be determined by examination of their gametes. For the three echinoderm species of interest eggs are yellow/orange in colour, while sperm is white. In addition, the sperm suspension is much more viscous than the egg suspension.

Eggs and sperm of both sea urchin species were obtained by vigorously shaking the animals until they shed gametes. Eggs were collected by placing the spawning female over a beaker with filtered sea water of appropriate salinity placed in ice, so that the animal would be partially submerged in the filtered sea water, aboral side up to make sure that the gametes are shed into the beaker with seawater. Appropriate salinity for echinoderm species: 37.8 parts per thousand (ppt) for *P. lividus* and 34.02 ppt for *S.purpuratus and P. miniata.* Sperm was collected by pipetting using a P200 micropipetter with appropriate pipette-tips from the surface of the spawning male into a 1.5 ml Eppendorf tube placed on ice. Sperm collected in such a way is referred to as dry sperm.

Gametes of sea star *P. miniata* were obtained by making a V-shaped surgical incision on the aboral side of the animal next to the gonads. The incision was then pried open carefully with thumb forceps, and another pair of forceps was used to collect part of the gonad from the incision. The female gonads were placed into Petri dishes on ice with filtered sea water of appropriate salinity. The male gonads were collected in the same way into a 1.5 ml Eppendorf tube placed on ice. Then the female gonads were torn up under the dissecting microscope using two pairs of thumb forceps to release oocytes into the sea water.

Eggs of all echinoderm species were then passed through a 200 µm nitex mesh into a 50 ml glass beaker to remove broken spines, tube feet, pieces of algae, that sea urchins feed on, and gonad debris (in case of *P. miniata*). *P. miniata* oocytes were also treated with 10 µM 1-methyladenine after passing through a filter to mature, until the germinal vesicle disappears. After germinal vesicle disappearance the eggs of *P. miniata* were washed with FSW and placed in a 50 ml glass beaker.

Prior to *in vitro* fertilization, excess water was removed from the beaker containing eggs, leaving a small amount of filtered sea water covering the eggs and 5 µl of dry sperm was diluted with 13 ml of filtered sea water of correct salinity in a 15 ml Falcon tube. Using a Pasteur pipette 10 to 20 drops of diluted sperm is added to the eggs. Volume of diluted sperm required to fertilize the eggs was dependent on the number of the eggs and the volume of water they were in. Fertilization was confirmed by the elevation of the vitelline membrane, visible under a dissecting microscope.

After fertilization, the embryos were cultured in 3 L of filtered sea water in 5 L glass beakers at 15°C for Pacific species (*S. purpuratus* and *P. miniata*), or at 18°C for Mediterranean species (*P. lividus*), until the embryos reach the required developmental stage.

2.2 Obtaining echinoderm embryo gut tissue

Sea urchin gut tissue was obtained by adapting existing protocols (McClay 2004; Juliano et al. 2014). Echinoderm embryos were grown as described above until

the desired stage. The embryos were then concentrated and collected into 1.5 mL Eppendorf tubes. The embryo suspensions were then centrifuged at 500 g for 5 minutes at 4°C to collect all embryos at the bottom of the tubes. The sea water was then removed and the embryos were washed once with 1 mL of Ca-Mg free sea water (31 g of NaCl, 0.8 g of KCl, 0.29 g of $NaHCO_3$ and 1.6 g of Na2SO4 in 1 liter of distilled water). After the wash the embryos are, again, centrifuged at 500 g for 5 minutes at 4°C to collect them at the bottom of the tube and treated with 1M glycine 0.02M EDTA in CaMg free sea water for 10 minutes on ice. After this incubation the embryos are pipetted up and down carefully using P1000 micropipetter three times and transferred onto 1% agarose plates under a dissecting microscope, to control the dissociation process. The gut tissue can be identified during this process as a small tube visible under the dissecting microscope. These tubes are then collected individually into a 1.5 ml Eppendorf tube with 100 µl of artificial sea water (28.3g NaCl, 0.77g KCl, 5.41g $MgCl_2 \cdot 6H_2O$, 3.42g $MgSO_4$ or 7.13g $MgSO_4 \cdot 7H_2O$, 0.2g $NaHCO_3$, 1.56g $CaCl_2 \cdot 2H_2O$ per 1 L of autoclaved milliQ water) placed on ice using a P10 micropipetter or a mouth pipette. Around 1000 individual guts were collected for each ATACseq and RNAseq sample.

2.3 ATAC-seq library preparation

ATAC-seq libraries were generated as described in Magri et al. Libraries were generated in collaboration with every author of the Magri et al: Marta Magri from Dr Jose Luis Gómez-Skarmeta lab, Jovana Ranđelović from Dr Giovanna Benvenuto lab, and Dr Claudia Cuomo from Dr Arnone lab. Cultured embryos

were collected by concentrating them using a 40 µm nitex mesh and washed on the mesh with artificial sea water. The concentrated embryos were then manually collected into 1.5 ml Eppendorf tubes with the aim to get 100000-135000 cells/nuclei in total. The embryos were then centrifuged at 500 g to remove liquid, and washed with artificial sea water twice in the 1.5 ml tube. The washing steps were necessary to remove all the remaining filtered sea water as the Mediterranean Sea water may have contaminants that may affect the downstream enzymatic reactions. The embryos were then resuspended in 50 µl of lysis buffer (10 mM Tris-HCl, pH 7.4, 10 mM NaCl, 3 mM MgCl2, 0.2% IGEPAL CA-630) and lysed by pipetting up and down for 3-5 minutes. Half of the lysate was used for counting released nuclei under a microscope using a haemocytometer with DAPI dye (1µl of 1:100 diluted DAPI in the 25 µl of the released nuclei). The other half was used to make up the tagmentation reaction by centrifuging the sample at 500 g, removing lysis buffer and then incubating for 30 minutes at 37°C with 25 µl of 2x tagmented DNA buffer (TD) (20 mM Tris-HCl, 10 mM MgCl2, 20% (vol/vol) dimethylformamide), 23.75 µl of nuclease free water and 1.25 µl of Tn5 enzyme. The Tn5 enzyme used for ATAC-seq library preparations was custom made, obtained from Dr Gómez-Skarmeta's, who collaborated on the ATAC-seq project.

After the reaction the tagmented DNA was purified using MinElute Kit (Qiagen) following manufacturer's instructions and eluted in 10µl of elution buffer. The eluted DNA was then amplified with a unique reverse primer (Buenrostro et al. 2013) to obtain the library for sequencing (10 µl of eluted tagmented DNA, 10 µl of nuclease free water, 2.5 µl 10 µM Nextera Primer 1, 2.5 µl 10 µM Nextera

Primer 2.X, where X is the unique barcode used for sequencing and 25 µl of NEBNext High-Fidelity 2x PCR Master Mix (New England BioLabs) using the following thermocycler program: 72°C for 5 minutes, 98°C for 30 seconds, then 15 cycles of 98°C for 10 seconds, 63°C for 30 seconds and 72°C for 1 minute, followed by a hold step at 4°C. The amplified library was then purified using MinElute Kit (Qiagen) following manufacturer's instructions and eluted in 20 µl of elution buffer. The quantity of the resulted amplified library was checked using Qubit dsDNA BR Assay Kit (Molecular Probes) and quality was assessed by running 70 µg of the library on a 2% agarose 1x TAE gel. A library was considered of good quality if two bands of ~200 bp and ~400 bp were visible. The bands correspond to single and two nucleosome DNA and spacer. Good quality libraries were then sent for sequencing.

2.4 Total RNA extraction for RNA-seq

Total RNA extraction was performed using RNAqueous-Micro Total RNA Isolation Kit (Invitrogen) according to the manufacturer's guidelines. Around 500 embryos or 1000 gut tissues were collected manually from the echinoderm cultures using a P10 micropipette into a 1.5 ml Eppendorf tube. Then the sample was centrifuged at top speed (16000 g) for 5 seconds. All liquid was removed, prior to adding 250 µl of lysis buffer, supplied with the kit, and vortexing thoroughly to ensure complete lysis. The RNA was then purified according to the kit instructions and eluted twice in 10 µl of nuclease free water to obtain 20µl total of eluted RNA per sample. Quality and quantity of resulting RNA was checked using Nanodrop ND-1000 (ThermoFisher Scientific) and Bioanalyzer with RNA 6000

Pico kit (Agilent). High quality RNA was sent for sequencing. Three biological replicates were collected per time point per condition (whole embryo or gut tissue only).

2.5 Tagging of putative CRMs

Tagging of putative CRMs was performed according to the protocol described by Nam et al. 2010 (Figure 2.1). Putative CRMs were selected by merging ATAC-seq peaks with less than 300 bp gaps from data from all available timepoints and 50 bp were added to the putative CRMs on both sides to ensure that the whole putative CRM is amplified from the genomic DNA as the primers were designed using Primer3web 4.1.0 (Untergasser et al. 2012) to fall within the added 50 bp. 18bp of the reverse complement of the beginning of the DNA tag sequence was added to the 5' of the reverse primer to ensure that the CRM can be combined with the DNA tag using overlap PCR (Xiong et al. 2006). Tag DNA was amplified from the MiniPrep of the Tag containing plasmids obtained from Dr Jongmin Nam. The amplified CRM and Tag sequences were purified using QIAquick PCR Purification Kit (Qiagen) and combined via overlap PCR using forward primer for the CRM and reverse primer (Table 2.1) for the Tag (Nam et al. 2010; Xiong et al. 2006) into CRM-Tag construct (Figure 2.1). The putative CRM and the chosen corresponding Tag were used in equal amounts to ensure efficient amplification via overlap PCR (Xiong et al. 2006). Expand High Fidelity PLUS PCR (Sigma) system was used for every PCR step. The resulting fragment was run on a 2% agarose 1x TAE gel, the required band was cut out, and gel-purified using GenElute Gel Extraction Kit (Sigma) according to manufacturer's guidelines and

eluted in 50 µl of Elution buffer from the kit. The eluted DNA was then re-purified using QIAquick PCR Purification Kit (Qiagen) and eluted in 30 µl of elution buffer from the kit. Yield and purity of DNA is assessed after each round of purification using NanoDrop ND-1000 (ThermoFisher Scientific). The second round of purification was performed to ensure purity of eluted DNA as it is likely that traces of agarose remain after gel extraction with GenElute Gel Extraction Kit (Sigma).

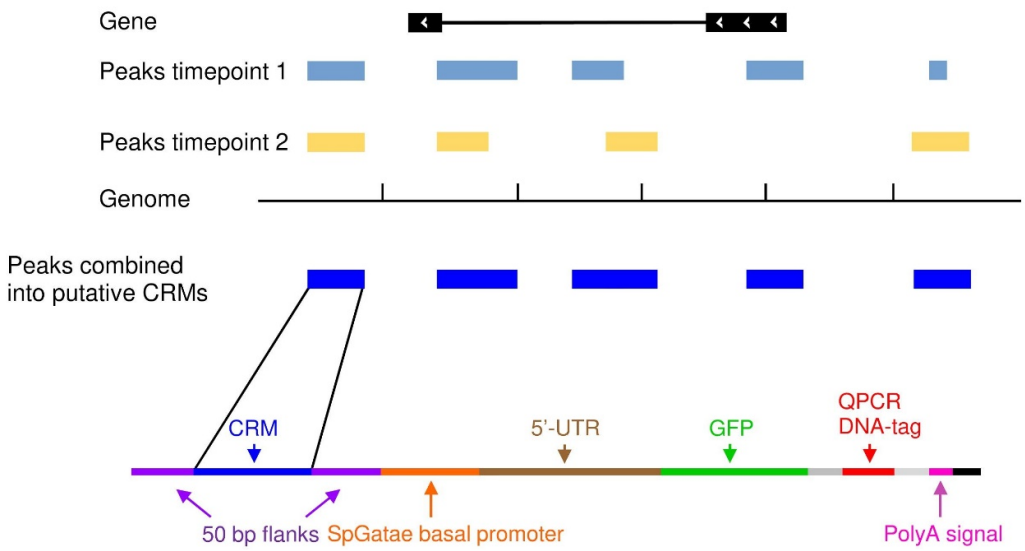


Figure 2.1 Schematic representation of CRM-Tag construct

Same procedure was performed for each putative CRM identified.

Table 2.1 Primers for amplification of CRMs from genomic DNA and for fusion with a Tag

CRM	Length	Forward primer	Reverse primer	Reverse CRM primer for fusion	Tag
SpCdx CRM1	679	TAACACATCTA ATGTCAT	TTTCAAATGGACGG GGATA	GAAGTAGCTGGCAGTGACGTGTTT CAAATGGACGGGGATA	Tag001
SpCdx CRM2	240	AGTTCAGAACA AATATTACCAG ACA	TGTCGAATTGCTTT ATTACGGA	AGTAGCTGGCAGTGACGTTGTCGA ATTGCTTTATTACGGA	Tag002
SpCdx CRM3	861	CGGGTTGGTTG ATTAGATGCA	TGTGGTCATCATTG GTCGAT	GAAGTAGCTGGCAGTGACGTTGTG GTCATCATTGGTCGAT	Tag003
SpCdx CRM4	280	CCCCAAACAT GAGTGCCAAA	AAAGGCGTTGGGGT GTTCTA	AGTAGCTGGCAGTGACGTAAAGGC GTTGGGGTGTTCTA	Tag004

SpCdx CRM5	336	GACACCAAAC CCAAACTCCC	TTGACAAAACCATT TAAGCAA	AGTAGCTGGCAGTGACGTTGACAA AACCATTTAAGCAA	Tag005
SpLox CRM1-4	1278	TCAATGCGGTG TCATGTGTT	ACCTTTAACCGGGG TCCT	AGTAGCTGGCAGTGACGTACCTTT AACCGGGGTCCT	Tag011
SpLox CRM5	409	GGCAACTAATA GCCGAGGTAT	TGGGCTGAATCGGG ATTTCT	AGTAGCTGGCAGTGACGTTGGGCT GAATCGGGATTTCT	Tag012
SpLox CRM6	222	CTTGATAAAAC AAATCCTCGGC A	GGAGAACCCGCCG GAAA	AGTAGCTGGCAGTGACGTGGAGA ACCCGCCGGAAA	Tag013
SpLox CRM7	159	CTTGACCGAAA CCGCGAG	AGGGTACTGGTGTT ACTTAGGA	AGTAGCTGGCAGTGACGTAGGGTA CTGGTGTTACTTAGGA	Tag014
SpLox CRM8	326	CCCTCTATCTC AATTCTAGAGA TCGT	TGTCTCAGAGCTAT ATTCAAAAACA	AGTAGCTGGCAGTGACGTTGTCTC AGAGCTATATTCAAAAACA	Tag015
SpLox CRM9	1082	TTCAGACGCCA TGGTGTAAA	TGTAAAATTGACAA AAAGACGATG	AGTAGCTGGCAGTGACGTTGTAAA ATTGACAAAAAGACGATG	Tag016
SpLox CRM9 Open	150	TAACACATCTA ATGTCAT	ATCACGTGTTGTCT TTTG	AGTAGCTGGCAGTGACGTATCACG TGTTGTCTTTTG	Tag015
SpFoxA_FI	393	GGCTGGTTGGT CACATGATC	GTACGTGCTCTTGG ATTGCC	AGTAGCTGGCAGTGACGTGTACGT GCTCTTGGATTGCC	Tag001
SpFoxA_J1	488	GCACTATTGGC CATGGGTTC	GTTTCTTAAGACTT GAGGGCCA	AGTAGCTGGCAGTGACGTGTTTCT TAAGACTTGAGGGCCA	Tag002
SpFoxA_F1	476	ACAAGACAAG AATAAACCATG CC	CCGAACTCCAATAA AATACATGT	AGTAGCTGGCAGTGACGTCCGAAC TCCAATAAAATACATGT	Tag003
SpFoxA_K1	364	AGACGATCTGT TCCCATACCA	TCCCTTCCCCAACA ATTTAACC	AGTAGCTGGCAGTGACGTTCCCTT CCCCAACAATTTAACC	Tag004
SpFoxA_JK	376	GCCTCATAAGC CTTCATGTCC	ACCCTTCAACGCCT GTATCA	AGTAGCTGGCAGTGACGTACCCTT CAACGCCTGTATCA	Tag005
SpFoxA_K2	373	GCAATTTAGCC AGAGACTTAA GG	AAAAGGGGAAACG GACGT	AGTAGCTGGCAGTGACGTAAAAGG GGAAACGGACGT	Tag006
SpFoxA_I1	728	GCCCATTCCAT TCACCCATT	TGACACATCTTCAT TCCCGAA	AGTAGCTGGCAGTGACGTTGACAC ATCTTCATTCCCGAA	Tag008

2.6 CRM Microinjections

The eluted DNA from CRM tagging was then used to make a pool of tagged CRM DNA for each gene of interest according to Nam et al. 2010. Microinjection

procedures were performed by Dr Maria I. Arnone. The resulting pool was used to make up microinjection solutions: 0.5 µl of tagged CRM pool, 1.2 µl of 1mM KCl, 0.275 µl carrier DNA (genomic DNA sheared with HindIII enzyme (2 units for µg of DNA, in SuRE/Cut Buffer B for 3 hours at 37°C), purified using QIAquick PCR Purification Kit (Qiagen) and diluted to 500 ng/µl) and the rest is water up to 10 µl (Arnone et al. 2004; Nam et al. 2010). The prepared solutions are then centrifuged at top speed for at least 15 minutes until microinjections. The eggs were dejellied in acidic seawater (pH 4.5) for 1 minute mechanically by pipetting, then washed in filtered sea water, and loaded in a mouth-pipette made over a flame to be wide enough for a single egg. The eggs were rowed onto protamine plates (treated with 4% protamine sulphate solution for exactly 1 minute and then washed with distilled water) filled with PABA-FSW (50 mg of p-aminobenzoic acid in 100 ml of filtered sea water of appropriate salinity) in a single file. The microinjecting needle, made from borosilicate glass with capillary by Sutter Instrument Co. Novato, CA pulled using P-97 micropipette puller (Sutter), was filled with the injection solution from the back using a Microloader pipette tip (Eppendorf). Loaded needle tip was broken off using a scratch in the middle of the protamine plate with eggs. The eggs were fertilized with a few drops of diluted dry sperm and injected with approximately 2 pL of microinjection solution. Injected eggs were then washed twice with filtered sea water and incubated at appropriate culturing temperature overnight. The next morning the hatched embryos were transferred to 4-well plates (ThermoFisher Scientific) with filtered sea water to grow until the desired stage.

2.7 Morpholino Perturbation Microinjections

Morpholino (MO) injections were performed using the same apparatus as in the CRM Microinjections section. The morpholino microinjection solutions contained 100 µM of final morpholino concentration. The co-injected CRM solutions were assembled as described in CRM microinjection section, adjusting added volume to 10 µl total volume taking into account the added MO solution. The morpholino containing microinjection solutions were warmed up to 75°C for 5 minutes and then passed through a 0.22 µm PVDF micro-filter (Millipore) placed in 500 µl tube by centrifugation at 2500 g for 2 minutes. After filtration the filter was disposed off and the filtrate was centrifuged for at least 15 minutes at top speed prior to microinjections. The microinjecting procedure was as described in the CRM Microinjections section. Morpholino sequence: SpHox11/13B translation morpholino – AAGCCTGTTCCATGCCGATCTGCA (Arenas-Mena et al. 2006).

2.8 Genomic DNA and mRNA extraction from CRM microinjected embryos

The microinjected embryos were grown to the selected stage and then collected into 1.5 ml tubes (Eppendorf). The collected samples were centrifuged at 500 g for 5 minutes to remove water leaving about 50 µl of water with the embryos at the bottom at the tube. If there is more than one tube per biological replicate then the remaining water and embryos of the same biological replicate were pooled together, centrifuged at top speed for 5 minutes to remove all the water and then vortexed in 350 µl of the RLT Buffer (with 2-mercaptoethanol added) from the AllPrep DNA/RNA Micro kit (Qiagen) to extract total RNA and DNA from these microinjected embryos. The lysate was loaded onto the DNA column from the kit

and centrifuged at 16000 g for 1 minute to bind DNA to the column. The column with DNA was transferred to another 2 ml tube, while the run-through lysate was mixed with 350 µl of 70% ethanol, loaded onto the RNA column from the kit and centrifuged at 16000 g for 1 minute to bind RNA. After centrifugation the run-through is discarded. The DNA column was washed once with 500 µl of AW1 buffer and twice with 500 µl of AW2, simultaneously the RNA column was washed once with 500 µl of RW1 buffer, once with 500 µl of RPE and once with 500 µl of 80% ethanol. The washed RNA and DNA columns were then centrifuged with the column lid open at top speed for 3 minutes to dry. After centrifugation the columns were placed into 1.5 ml Eppendorf tubes for elution. DNA was eluted in 100 µl of 65°C nuclease free water while RNA was eluted in 18 µl of 65°C nuclease free water. The RNA was then treated with DNase (1 µl of 2U/µl DNase and 2 µl of DNAse I buffer from RNAqueous-Micro Total RNA Isolation Kit (Invitrogen) added to the full eluted RNA volume) for 30 minutes at 37°C, post-treatment DNAse was deactivated by adding 2 µl of DNase Inactivation Reagent from the same RNAqueous kit, incubating for 2 minutes at room temperature and vortexing twice: once at the beginning of the incubation and after 1 minute. The eluted and DNase treated RNA was then removed from the DNase Inactivation Reagent by centrifuging at top speed for 30 seconds and transferring 14 µl of the supernatant to a 200 µl PCR tube to synthesize cDNA from the RNA using SuperScript VILO cDNA Synthesis Kit (Invitrogen) (14 µl of eluted DNA, 4µl of 5X VILO Reaction Mix and 2 µl of 10X SuperScript III Enzyme Blend). The synthesized cDNA was then used for qPCR quantification.

2.9 qPCR quantification of CRM expression

Extracted genomic DNA and synthesized cDNA were used to estimate relative expression of each microinjected CRM-Tag construct. Prior to the qPCR the cDNA was amplified using universal primers (Nam et al. 2010) due to low amount of cDNA synthesised from extracted mRNA using the following thermocycler program: 2 minutes at 95°C, 21 cycles of 15 seconds at 95°C, 30 seconds at 60°C and 1 minute at 72°C, followed by 5 minutes at 72°C and then a hold step at 4°C indefinitely. The product was purified using QIAquick PCR Purification Kit (Qiagen) and eluted in 30 µl. The elution was then used for qPCR quantification. The reactions were assembled to contain: 5 µl of Fast SYBR Green Master Mix, 4 µl of qPCR primers at 0.7 pmol/ul each and 1 µl of cDNA/gDNA. The quantification was performed in Life Technologies ViiA7 Real–Time PCR System machine using the following program: 20 seconds at 95°C, then 40 cycles of 1 second at 95°C and 20 seconds at 60°C, followed by melting curve stage at 95°C for 15 seconds, 60 °C for 1 minute and 95°C for 15 seconds again. After the run, the results were collected on a USB stick and exported to a table format. In order to determine the relative construct expression levels total GFP was used as control for both cDNA and genomic samples containing same specific tag primers (Table 2.2). The number of tags expressed was normalized to number of tags incorporated into genomic DNA by dividing number of expressed tags in cDNA by number of expressed tags in gDNA relative to GFP. At least two biological replicates of qPCR tag expression quantification were performed per injected CRM pool (up to 7 CRM-Tags per microinjected pool) per time point. Values obtained were averaged between replicates for plotting. Whether the

microinjected pool contains active CRM-Tag constructs was determined by visualization of GFP expression under the microscope.

Table 2.2 Tag qPCR primers

Tag	Length	Forward Primer	Reverse Primer
Tag001_QPCR	187	ACCACGTGTCCAGTGTGTGTG	AAGGTGGCGGTTTCGCCTCTA
Tag002_QPCR	187	ACGAAGCTGGTAGAGTGCTGG	GTCCCGCTTTAAGACGGTGAG
Tag003_QPCR	183	CACGACATCCGTAAGCCCA	TGTCCTACGTGCACAAGCA
Tag004_QPCR	183	TGGATCTGCCGACAACCAG	GGCTTCAAGGACCGATCAC
Tag005_QPCR	187	GTTGCGTTCCAAACGTCGTGG	CCTGGGGTATGTCGCGTATCA
Tag006_QPCR	186	GTCGCATCTTGCCAGTTGG	AGTCCGCATTACACATGCGACG
Tag007_QPCR	187	AGCTGAAACAAGGATTGCGGTG	ACCGCTCACTAGCTGAGACG
Tag008_QPCR	186	TCGCTATCACTGACGCGAG	CAAAGGAACCAAGCGAATCCTG
Tag009_QPCR	186	TGTGTCGTAGTTCCACCGA	GACTGTTTAGAGGGCGTTTGAC
Tag010_QPCR	186	CCAAGATCAGCGACATGGTC	GGATTCGAAAGTTGTCACCCA
Tag011_QPCR	186	TACTCGTCCGGCGTCACAA	CTTATGTCGGCACGGAATGACC
Tag012_QPCR	184	GTGCACTTCGTGTGTGCGTG	TCTGAACCACACGGTGGAA
Tag013_QPCR	184	CAGAACACCGAGGCACCAA	ATACGGCTGTTCGACAGGAA
Tag014_QPCR	186	CCTACAGAAGTTCACAGGTCCA	TACGCGACTGCGATGAGAC

Tag015_QPCR	**187**	**CAAATCCTGCAAGCGCGGAA**	**GAGTATTCGCTACATCCAGCCA**
Tag016_QPCR	**187**	**TTGCGAGCGCCGAACTGGTTT**	**GTGCATGCATTGAGTCTCGGC**
Tag017_QPCR	**188**	**ACCAAGTCAGTGATTGCGCGA**	**CAAGGTCACTGTGTGGTGTTCG**
Tag018_QPCR	**188**	**TTGCAGTAATTCACGAGGCCAA**	**TTCGCGATCTGTCCACAACGA**
Tag019_QPCR	**186**	**CAATGGGTGTATTGTGGGTTGC**	**CCGTAAGGTCGCTCCAGTA**
Tag020_QPCR	**185**	**CACTTGACGTGCTTGGAAGC**	**CCTCCAATTCTGGCAGACAC**
TagAmplification	**2007 to 2012 bp total TAG length**	**ACGTCACTGCCAGCTACTTC**	**TAATACGACTCACTATAGGG**
TagFusion	**Total TAG length plus CRM length**		**CACAAACCACAACTAGAATGCA**

2.10 CRM expression visualization

The microinjected embryos are grown to a selected stage and then collected either onto a two-well slide or a glass bottom Petri dish. The two-well slides were used for visualization using Zeiss Imager.Z2, while the glass bottom Petri dish were used with the confocal Zeis LSM-700. A few drops of 100% methanol were added to the collected embryos to prevent them from moving, and, unlike other fixatives, methanol does not destroy the GFP fluorescence. The CRM driving GFP expression images were taken using the GFP wavelength light-source, and a bright-field image when possible to show the localization of the driven GFP expression (Arnone et al. 2004). These images were then combined into a single image using (Fiji is just) ImageJ 1.52o.

2.11 ATAC-seq data mapping

The sequencing results in FASTQ format were checked using fastqc 0.11.5 (Andrews 2010). The reads were of good quality so no trimming was performed. Reads were mapped to the corresponding genomes: *S. purpuratus* 3.1 genome (Sea Urchin Genome Sequencing Consortium et al. 2006) and *P. miniata* 2.0 genome from Echinobase.org (Cary et al. 2018; Kudtarkar & Cameron 2017), *P. lividus* genome was obtained as part of sequencing consortium) using bowtie2 2.3.4.1 (Langmead & Salzberg 2012) (Table 2.3). The read alignment was adjusted with +4 bp offset on the positive strand and with -5 bp on the negative strand (Magri et al. in press). The resulting fragments in BAM format were filtered to keep only fragments of less than 130 base-pairs (nucleosome free regions) and then converted into bed files using bedtools 2.27.1 (Quinlan & Hall 2010). The resulting files are then fed to MACS2 2.1.2 software (Zhang et al. 2008) to call peaks using BED as input file format as well as setting extsize to 100, shift to 50 and using nomodel setting to prevent model building; corresponding genome sizes were also specified for each species in MACS2: 936564995 for *S. purpuratus,* 927475755 for *P.lividus* and 1006316195 for *P. miniata* in base-pairs (Table 2.3); all the other MACS2 settings were kept at default. In order to combine replicates for further analysis bedtools intersect tool (Quinlan & Hall 2010) was used with default settings using A replicate as file a and B replicate as file b. Bedtools intersect was used to ensure higher stringency of replicate combination, so that only regions present in both replicates would end up in the combined file. The resulting files were used in subsequent analyses.

Table 2.3 Table of genomic resources used for mapping chromatin accessibility and transcriptomic datasets. Effective size is the size of the genome assembly (in bp) omitting the Ns.

Species	***S. purpuratus***	***S. purpuratus***
Assembly version	**v3.1**	**v5.0**
Number of scaffolds	32008	808
Size (in bp)	936564995	844507448
Effective Size (in bp)	815936258	844170948
Scaffold N50 (in megabases)	0.4016	37.3
Number of genes	29072	29033
Genome file	Spur_3.1.LinearScaffold.fa	spur5.fasta
Annotation Files	Transcriptome.gtf	SPU_ids_on_spur5.gff3
Reference	Kudtarkar & Cameron 2017; Tu et al. 2012	*S. purpuratus* sequencing consortium
Species	***P. lividus***	***P. miniata***
Assembly version	**no version number**	**v2.0**
Number of scaffolds	3747	57698
Size (in bp)	927475755	1006316195
Effective Size (in bp)	879641317	990147015
Scaffold N50 (in megabases)	41.5	0.0763
Number of genes	30593	30399
Genome file	Pliv_PqN3S_sm.fa	pmin_scaffolds_v2.0.fa
Annotation Files	Pliv_PqN3S_evmp.gtf	genes_Pm.gtf
Reference	*P. lividus* sequencing consortium	Kudtarkar & Cameron 2017
Species	***B. lanceolatum***	
Assembly version	**no version number**	
Number of scaffolds	10247	
Size (in bp)	495353434	
Effective Size (in bp)	474928346	
Scaffold N50 (in megabases)	1.29	
Number of genes	21428	
Genome file	Bl71nemr.fa	
Annotation Files	Bl_Annotation.gtf; gene_models_only_BraLan.gff3	
Reference	Marlétaz et al. 2018	

2.12 Differential ATAC-seq analysis

Bedtools merge was used to combine peaks from different conditions (whole embryo and gut only) of *S. purpuratus* ATAC-seq data into one file, then bedtools

makewindows was used with window size of 150 to split the peaks into small regions, approximately one nucleosome in size, to maximize resolution of differential analysis. FASTA sequences of these regions were obtained using the resulting file and bedtools getfasta tool with default settings using the *S. purpuratus 3.1* genome (Sea Urchin Genome Sequencing Consortium et al. 2006). Then the ATAC-seq raw FASTQ reads from different conditions were pseudoaligned onto these FASTA sequences using Salmon 0.11.3 (Patro et al. 2017) to generate a count table for each of the regions in every condition and biological replicate. The resulting counts table was then used in DESeq2 1.22.2 R package (Love et al. 2014) to perform differential analysis comparing condition (e.g. gut data) with control (e.g. whole embryo data). RUV 1.16.1 package was used in addition to DESeq2 to remove sources of unwanted variation (Risso et al. 2014). Regions that are differentially more accessible in the condition, e.g. gut, compared to control, e.g. whole embryo, with the p-value of less than 0.1 were treated as significant and were saved for visualization in genome browsers and for further analysis. P-value cut-off of 0.1 was used since ATAC-seq data presents an indication to where a region, which could be a functional CRM for a particular tissue type, is; differential ATAC-seq data allows to narrow down the search by examining these differential windows first. These windows need to be tested and validated *in vivo,* and, as such, potential false positives are less of an issue in differential ATAC-seq analysis compared to RNA-seq differential analysis, which is supported by using an arbitrary p-value for differentially accessible locations by Shashikant and Ettensohn (Shashikant & Ettensohn 2019). HOMER *de novo* motif discovery tool findMotifsGenome.pl (Heinz et al.

2010) was used to predict, which transcription factors could be bound to these regions.

2.13 *In silico* GRN drafting

The peaks from MACS2 peak calling and from differential ATAC-seq analysis can be used for analysis in HOMER 4.10.3 (Heinz et al. 2010) to increase GRN resolution *in silico*, as described in Lowe et al. 2019. The current project concerns open chromatin regions around ParaHox genes and thus the protocol described in part IV A ii of the protocol paper was used for GRN drafting in the context of this project (Lowe et al. 2019). The protocol described in the paper was developed in collaboration with Dr Elijah K. Lowe and Dr Claudia Cuomo. In order to identify which transcription factors could be affecting ParaHox genes, putative CRMs were identified by merging ATAC-seq peaks with less than 300 bp gaps from data from all available timepoints and then these putative CRMs were annotated with the nearest gene using the annotatePeaks.pl tool from the HOMER package. Then only the putative CRMs identified as close to ParaHox genes were selected for all ParaHox genes together and for each gene separately. Putative transcription factor motifs were identified in these select CRMs using HOMER (Heinz et al. 2010) and JASPAR2018 database (Khan et al. 2018). The JASPAR matrices files were converted to HOMER-usable motif files through the use of a custom script (Siebert et al. 2018). At this stage, transcription factors can be filtered using tissue specific data, e.g. keeping only the TFs that are found in the digestive system of the sea urchin using tissue specific transcriptomic data. Then putative CRMs near these TFs were also

selected and motifs of the identified and filtered transcription factors were searched for in these selected CRMs as well. This allows to make an *in silico* draft of an interconnected gene-regulatory network around ParaHox genes (Lowe et al. 2019), nodes and interactions of which need to be validated *in vivo* (Figure 2.2).

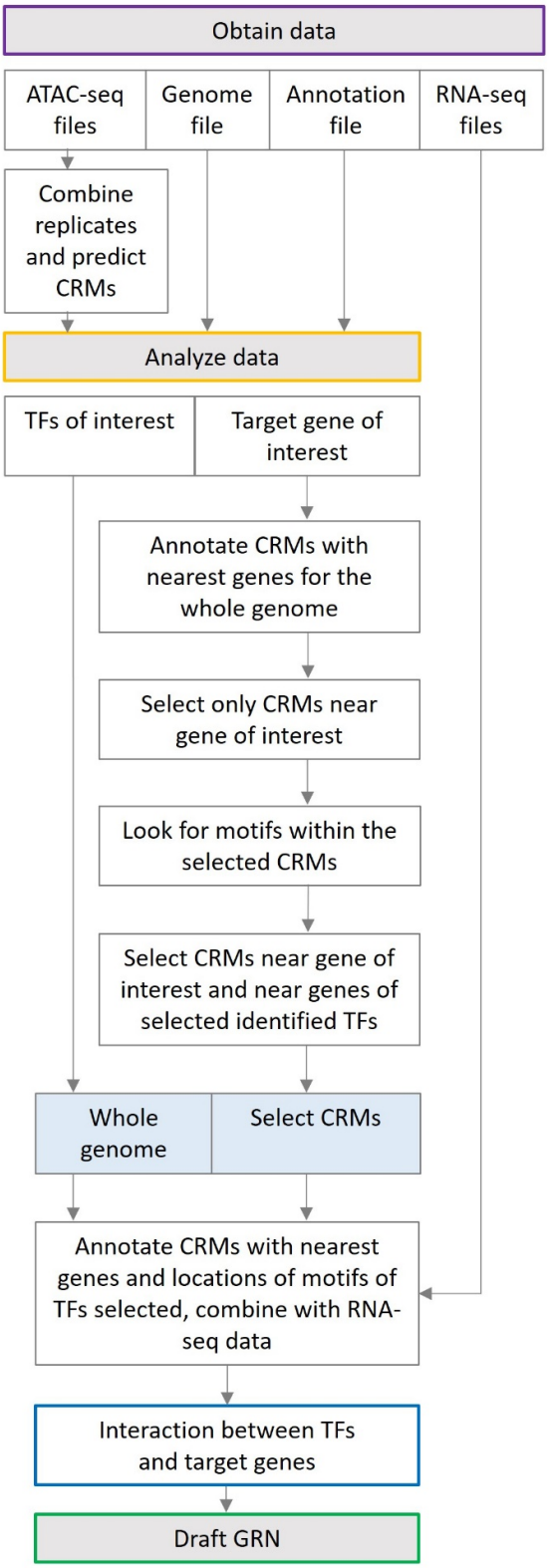


Figure 2.2 Flowchart detailing steps of the *in silico* GRN drafting approach. Adapted from Lowe et al. 2018.

2.14 PCA motif analysis

In order to perform evolutionary comparisons into the wiring of the GRN around ParaHox genes, a motif occurrence principal component analysis (PCA) was performed. The transcription factor motifs were identified as described in the previous section of Materials and Methods of this thesis. Resulting lists of potential TF motifs and their counts were combined in a counts-like table, which was then fed into DESeq2 to perform PCA analysis on presence of motifs among the selected species: *S. purpuratus*, *P. lividus*, *P. miniata* and *B. lanceolatum* (Figure 2.3). ATAC-seq peaks for *B. lanceolatum* were obtained from Dr José Luis Gómez Skarmeta. DESeq2 PCA analysis involved estimating size factors and dispersions and using Wald test for the GLM coefficients since the count differences were less two orders of magnitude (Love et al. 2014). DESeq2 (Love et al. 2014) PCA plots and Limma (Ritchie et al. 2015) Venn diagrams were used for visualization of this analysis.

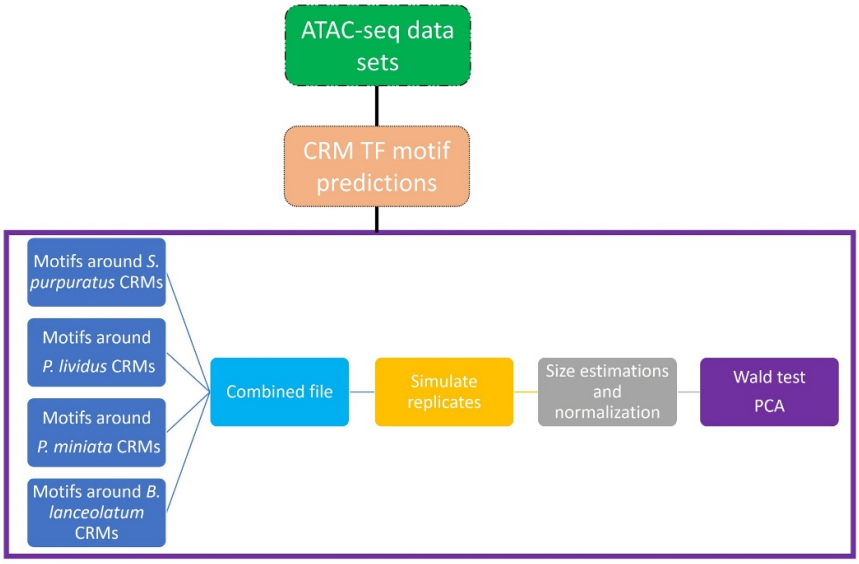


Figure 2.3 Flowchart of transcription factor motif count PCA analysis

2.15 RNA-seq data mapping

Sequenced RNA-seq data was quality checked using fastqc 0.11.5 (Andrews 2010), and bad quality sequences were trimmed from the reads using Trimmomatic (Bolger et al. 2014). The settings used were paired-end, phren33 quality scores, as well as using the software-provided TruSeq3-PE-2-mod.fa file to remove sequencing adapters using ILLUMINACLIP setting to allowing 2 mismatches and keeping only bases with quality over 30 for paired end reads. In addition, to remove bad quality reads a SLIDINGWINDOW was used with width of 3 bases, which cut when the average read quality fell below 25. In the end, only the reads that are at least 25 bp long were kept after trimming. The paired output of Trimmomatic was used as input for Salmon, ran for unstranded paired end read library with reads facing each other with other settings left as default. Salmon index for *S. purpuratus* was made using Transcriptome fasta file

(corresponding to Transcriptome.gtf annotation file shown in Table 2.3) from Echinobase.org (Cary et al. 2018; Tu et al. 2012) with the k value of 25. The resulting quants files were loaded into DESeq2 (Love et al. 2014) R package for differential expression analysis.

2.16 Differential expression analysis

Differential expression analysis was performed using DESeq2 (Love et al. 2014) package using the three biological replicates as stated in Total RNA extraction for RNA-seq section. The analysis was performed using only condition as the factor since it was impossible to keep the batches for whole embryo or gut samples consistent: whole embryo RNA was collected from the cultures prepared for ATAC-seq to allow comparison with ATAC-seq data, due to labour intensiveness the gut samples for RNA-seq were collected separately. RUVSeq (Risso et al. 2014) with k (number of sources of variability) of 2 was used to remove sources of unwanted variation treating transcripts with the p-value of more than 0.1 as non-significant for the RUV analysis. The RUVSeq-called variable factors were used in as part of design option for DESeq function to account for these factors during differential analysis of whole embryo and gut enriched RNAseq data. The results were deemed significant if their p-adjusted value was less than 0.05. Differential expression analysis results were saved as a csv file and annotated, adding *S. purpuratus* gene names.

2.17 HiC data analysis

Raw HiC fastq files for 50 hpf *S. purpuratus* were obtained from Dr José Luis Gómez Skarmeta. To perform the HiC analysis a Bowtie2 index of the *S. purpuratus* genome version 5 (*S. purpuratus* sequencing consortium, personal communication) (Table 2.3) was build using the default settings. The HiC library was prepared using DpnII restriction enzyme, thus restriction fragments were generated using digest_genome.py script from HiC-Pro 2.11.1 package (Servant et al. 2015). HiC-Pro was used to analyse HiC data and obtain contact matrices for the 21 chromosome sized scaffolds of *S. purpuratus* genome version 5 at 100 kb resolution and for ParaHox genes loci at 20kb resolution with the whole genome. The resulting matrices were used for further analysis in HiTC 1.26.0 (Servant et al. 2012) R package to analyse interactions and visualize TADs for the loci of interest. BioCircos 0.3.4 (Cui et al. 2016) was used for visualization of physical contacts between ParaHox genes. HOMER (Heinz et al. 2010) was used to assess the significance of the contacts identified.

2.18 Single Cell RNA-seq data analysis

Single cell RNAseq library preparations were performed by Periklis Paganos and then sequenced in collaboration with Dr Detlev Arendt's lab. Therefore the BCL read files of the scRNAseq sequencing runs were converted into FASTQ by Dr Jacob M. Musser using Cellranger 3.0.2 (10xgenomics), a postdoc from Dr Arendt's lab, so the FASTQ files were downloaded from him. Four biological replicates of *S. purpuratus* pluteus at 72hpf (72 hpf 1, 72 hpf 2, 72 hpf 3, 72 hpf 4) and two technical replicates for two of these biological replicates (72 hpf 1 E,

72 hpf 2 E, which were constitute extra sequencing of samples 72 hpf 1 and 72 hpf 2). Samples 72 hpf 1, 72 hpf 2 were obtained using Single Cell 3' 10xgenomics v2 chemistry sets, while samples 72 hpf 3 and 72 hpf 4 were obtained using v3 chemistry sets. The genomic index for scRNA-seq data for *S. purpuratus* was made via cellranger mkref command using the genome version 3.1 (Sea Urchin Genome Sequencing Consortium et al. 2006; Kudtarkar & Cameron 2017), due to it having the most complete annotation at the time of analysis and writing, and using the annotation file which can be found on Echinobase (Tu et al. 2012; Kudtarkar & Cameron 2017) and converting it from gff3 format to gtf format using gffread tool of the cufflinks suite (Trapnell et al. 2010) in order to make it compatible with CellRanger (10xgenomics). *SpLox* is present in the genome version 3.1 twice on two scaffolds, however it is known that there is only one *SpLox* gene in sea urchin (Arnone et al. 2006), so the shorter transcript (sequence exactly equal to an exon of actual *SpLox* gene) was removed from this annotation file, this was done for scRNA-seq analysis and not for other analyses since read mapping in scRNA-seq analysis affects cell clustering and having a extra *SpLox* can lead to wrong cell clustering. The downloaded data was mapped and a count matrix was generated using CellRanger 3.0.2 (10xgenomics). The cell number forcing was estimated by running the cellranger count command without forcing to estimate cell numbers and then checking the web summary file and re-running cellranger count with the chosen forced cells number. Samples 72 hpf 1 and 72 hpf 1 E were mapped forcing 4000 cells, 72 hpf 2 forcing 7000 cells, while samples 72 hpf 2 E, 72 hpf 3 and 72 hpf 4 were all mapped with 5000 cells setting. The resulting count

matrices were used for further analysis in Seurat 3.0.2 package in R (Stuart et al. 2019). The data was loaded as Seurat objects filtering out genes that are transcribed in less than three cells and cells that have less than certain number transcribed genes, this number was selected based on the feature scatter plots. The objects were renamed and added into an R list, the datasets in the list were normalized, variable features were found using vst method and setting top variable features to 2000, then the anchors were found for integration of the six objects in the list into one for downstream analysis. The data in the integrated object was scaled and principal component analysis (PCA) was performed using variable features of this single Seurat object. Shared Nearest Neighbor (SNN) graph was computed with 20 dimensions to find resolution one clusters in the next step. The clustering was dimensionally reduced using Uniform Manifold Approximate and Projection (UMAP) in the Seurat package using 20 dimensions. The number of dimensions was selected by plotting significant dimensions using the JackStraw plots and standard deviations using the elbow-plot for the PCA. Then positive cluster markers were identified of the RNA assay of the data using the genes that are detected in at least 0.01 fraction of min.pct cells in the two clusters. The cluster markers were used to rename clusters according to their putative identity. The resulting table was then annotated using Linux command-line tools adding PFAM terms (Trapnell et al. 2010; Finn et al. 2014) for associated proteins, gene-ontology terms and other descriptions from Echinobase (Kudtarkar & Cameron 2017). Information on the expression of each mapped transcript in every cluster was obtained by converting a Seurat DotPlot with all these transcripts as features into a table using ggplot_build from the

ggplot2 3.2.0 R package (Wickham 2016). Transcription factors with Average Expression of at least 0.5 were then extracted for each relevant gut cluster from this table.

2.19 Obtaining gene information

In order to gain greater insight into the function of the genes and identify transcription factors, a number of protein BLAST searches was performed using NCBI-BLAST+ 2.7.1 (Altschul et al. 1990) via using the *P. lividus*, *P. miniata* and *B. lanceolatum* predicted proteins as queries and *S. purpuratus* proteins from Echinobase (Kudtarkar & Cameron 2017) and human proteome from UniProt (Apweiler et al. 2004) as BLAST databases keeping only the first hit. In addition to these, *S. purpuratus* has a wide range of information available at Echinobase.org such as functional annotation, gene ontology terms and various associated IDs (WHL, GLEAN etc). To complement these, PFAM terms for *S. purpuratus* proteome were identified using hmmscan from hmmer 3.1.2 (hmmer.org) (Eddy 2011) and PFAM-A HMM library (Finn et al. 2014) on SPU_peptide.fasta (Kudtarkar & Cameron 2017) and then filtering the results to have the E-value of less than 0.00005. All these were modified to have one entry per WHL ID, since this ID is used in the annotation gff3/gtf file and is also present in the transcriptome FASTA file (Tu et al. 2012; Kudtarkar & Cameron 2017).

2.20 *S. purpuratus* genome version 5.0 annotation

The *S. purpuratus* genome version 3.1 genes were mapped onto the *S. purpuratus* genome version 5.0. In order to achieve this, a GMAP index of the

new genome was generated using GMAP version 2017-11-15 (Wu & Watanabe 2005) with default settings. SPU_Nucleotides.fasta file (Kudtarkar & Cameron 2017) was used for mapping onto the new genome using the generated index with default settings and gff3_gene output format. Resulting annotation was then visualized using Integrative Genomics Viewer (IGV) 2.4.1 (Robinson et al. 2011).

2.21 Putative CRM visualization and sequence similarity assessment

Predicted putative CRMs were loaded into Integrative Genomics Viewer (IGV) for visualization along with relevant ATAC-seq peaks and gene annotations, and other information such as transcription factor motif location and, in case of the sea urchin species, similar regions of the predicted CRMs between species for the genes of interest. Putative ParaHox CRM sequences from each species of interest were extracted using bedtools getfasta tool (Quinlan & Hall 2010) from their respective genomic sequences. The resulting FASTA files were compared in a pair-wise manner using BLASTN search, sequences with raw score above 100 were kept. Locations of similar sequences on the genome of the species they were searched against were extracted using IGV Find Motif tool (Robinson et al. 2011) and then visualized using the same genomic browser.

2.22 Statistical analysis

Majority of the tools used to perfrom the analyses, results of which are presented in this thesis, have statisitical methods associated with them. However, some of the analysis performed was not statistically tested since it was not performed using specialized tools. This kind of analysis includes ATAC-seq peak distribution

in relation to gene model in the four species, discovery of known CRMs in the obtained ATAC-seq peak sets for *S. purpuratus*, indentification of differentially more open chromatin regions near the differentially expressed genes from the samples of the gut tissue, and, finally, morpholino injections and its effect on CRM activity. Statistical analysis was performed for these results.

2.22.1 Statistical testing of ATAC-seq peak distribution

In order to determine whether the obtained peak distributions in the four species are not random and, likely, have an underlying biological meaning, a random peak set was generated for each species using bedtools shuffle tool (Quinlan & Hall 2010), using the corresponding putative CRM sets for each species as input (see section 2.13). Thus random set has the sizes from the putative CRM set and the total number of random peaks is equal to number of putative CRMs. Then these random CRMs were annotated using HOMER annotatePeaks.pl tool (Heinz et al. 2010) to obtain the distribution of random CRMs in relation to gene annotation features, such as promoter and transcription start sites (TSS), exons, introns, transcription termination sites (TTS), or intergenic regions. The number of random CRMs in a species per each feature was divided by total number of random CRMs and multiplied by total number of ATAC-seq peaks of the time point, to make sure that the observed distrubition at a given time point is comparable to the expected distribution (obtained using random CRMs) and that they have the same total counts. The counts per each gene annotation feature were used for chi-squared test using Python 3.7.3 SciPy 1.2.1 chisquare tool to obtain the chi-squared statistic and the associated p-value. In addition, a separate test was performed for promoter-TSS peaks to determine whether the

number of peaks observed in this feature is significant. The significance of the number of genes with proximal peaks was also tested via Python 3.7.3 SciPy 1.2.1 chisquare using the same random CRM sets.

2.22.2 Statistical testing of overlap between known CRMs with ATAC-seq peaks

The random CRM set for *S. purpuratus* described in 2.22.1 was also used to determine whether the known CRMs can be found in the ATAC-seq peaks by random. The numbers of overlapping and non-ovelapping known CRMs with the actual putative CRM set were used as the observed values for Python 3.7.3 SciPy 1.2.1 chisquare test, while the numbers of the overlapping and non-overlapping known CRMs with the random CRM set were used as expected counts. This test was used to show the significance of the overlap between known CRMs and actual ATAC-seq peaks.

2.22.3 Statistical testing of finding differentially open peaks near differentially expressed genes

In order to determine whether the same number of peaks, differentially more open in the gut compared to whole embryo, located near the genes that are differentially more expressed in the gut, could be obtained by random, the differentially more open peaks were shuffled within the locations of all putative CRMs using bedtools shuffle (Quinlan & Hall 2010) to obtain the random 150 bp peak set. These small random peaks were annotated using annotatePeaks.pl tool (Heinz et al. 2010) to find nearest genes. The number of genes that was identified through differential RNA-seq analysis as ones expressed greater in the gut sample that were also near the random 150 bp peaks were treated as expected

random values for the Python 3.7.3 SciPy 1.2.1 chisquare test. Number of genes near actual differentially more open in the gut chromatin regions was used as the observed values, to show that the number of genes identified with differentially more open peaks is not obtained randomly.

2.22.4 Statistical testing of SpHox11/13B morpholino on SpLoxCRM9

In order to determine whether the effect of SpHox11/13B morpholino (MO) on SpLoxCRM9 expression is significant the chi-squared test using Python 3.7.3 SciPy 1.2.1 chisquare was performed. The total numbers of MO and SpLoxCRM9 injected embryos expressing and not expressing GFP were used as observed values, the total numbers of control only SpLoxCRM9 injected embryos expressing and not expressing GFP were used as expected values to obtain the chi-squared statistic and associated p-value and to show the significance of MO effect.

Contribution Statement

Dr Maria I. Arnone has performed the microinjection procedures. Author of this thesis has prepared the required solutions, such as PABA-FSW, artificial sea water and microinjection solutions, and equipment, such as protamine coated plates and microinjection needles.

Dr Elijah Kareem Lowe and Dr Claudia Cuomo, at the time of collaboration from Dr Maria I. Arnone lab, have collaborated on in silico GRN drafting protocol establishment. Dr Lowe has devised the approach, Dr Cuomo has contributed the RNA-seq protocol for sea urchin species. The author of this thesis, Danila

Voronov, has tested the approach described in the protocol and expanded it with the use of scRNA-seq data.

Marta Silvia Magri from Dr Jose Luis Gómez-Skarmeta lab, Jovana Ranđelović from Dr Giovanna Benvenuto lab and Dr Claudia Cuomo from Dr Maria I. Arnone lab have collaborated on ATAC-seq library preparation. Jovana Ranđelović has contributed to preparation of libraries for the sea urchin species, including gut separation. Dr Claudia Cuomo from Dr Maria I. Arnone lab have collaborated on ATAC-seq library preparations for *P. miniata*. Marta Magri, in addition to contributing to library preparations for every species, also collaborated on initial data analysis: read mapping and peak calling. Author of this thesis has contributed to preparation of libraries for every echinoderm species described, as well as performed data analysis, results of which are described in chapters 3, 4 and 5 of this thesis.

Periklis Paganos from Dr Maria I. Arnone lab has performed single cell RNA-seq library preparation and fluorescent in situ hybridizations on scRNA-seq cluster markers. Dr Jacob Musser, at the time of collaboration from Dr Detlev Arendt lab, performed scRNA-seq sequencing and collaborated on data analysis. Author of this thesis has contributed to whole pipeline of analysis of the scRNA-seq data, post FASTQ file generation.

S. purpuratus sequencing consortium resulted in new genome assembly (version 5.0). Author of this thesis has performed mapping of existing gene sequences from *S. purpuratus* assembly 3.1 onto the genome assembly 5.0 in order to

assess the quality of the assembly. Results of this mapping are described in the context of ParaHox genes in the chapter 3.

Other work described and presented in this thesis was performed by the author of this thesis.

CHAPTER 3

PARAHOX CHROMATIN ORGANIZATION ASSESSED BY GENOME ASSEMBLIES, ATAC-SEQ AND HIC DATA

This chapter contains results pertaining to the organization of the ParaHox genes in the nucleus of the four deuterostome species: their linear chromosomal locations, three-dimensional organization of the ParaHox genes in a sea urchin and chromatin accessibility around these genes. Detailed results discussed in this chapter can be found in the Non-book component files on the USB drive.

3.1 Introduction

As stated in chapter 1 chromatin organization and access plays a crucial role in gene expression regulation. Thus, part of this thesis was to gain insight into the organization of ParaHox genes on the chromosome, especially for the sea urchin species, for which actual ParaHox gene locations are not clear from the published genome data (Sea Urchin Genome Sequencing Consortium et al. 2006; Kudtarkar & Cameron 2017; Cary et al. 2018), and to determine whether, despite lacking a tight ParaHox cluster, the ParaHox genes in sea urchin might still be located close to each other in the three-dimensional organization of the chromatin, as well as to assess the chromatin accessibility around the ParaHox genes to predict putative CRMs.

To address the question of the chromosomal localization of the ParaHox genes, high-quality genome assemblies are necessary. Ideally, after the assembly each chromosome present in the haploid nucleus will be represented as a single scaffold. However, due to repetitive sequences, such as GC-rich stretches, microsatellites and tandem repeats, and insufficient length or quality of read complete chromosomes cannot be assembled into single scaffolds. Therefore, a chromosome in the genome is frequently represented by multiple scaffolds when it is impossible to deduce whether these scaffolds are contiguous. This has led to the issue observed in the *S. purpuratus* genome assembly version 3.1: the three ParaHox genes are on different scaffolds. *SpGsx* is located on Scaffold550 that spans across 433 kilobases (kb), *SpLox* is on Scaffold1640 which is 445 kb long, *SpCdx* is on Scaffold663 which is 427 kb in size. There is no information

available to the contiguity of these scaffolds, so it is also impossible to gain insight into relative locations and orientations of these genes in the genome, but the sizes of the scaffolds suggest that there is no intact ParaHox cluster in *S. purpurarus*. Similar issue exists in the *P. miniata* genome assembly where the ParaHox genes occupy two different scaffolds. *PmGsx* is located on Scaffold5544 spanning 247 kb, while *PmLox* and *PmCdx* are on the same scaffold, Scaffold4298 42 kb in size. However, due to the genomic proximity of these three genes they were identified to be in a cluster by BAC sequencing (Annunziata et al. 2013; Kudtarkar & Cameron 2017), which makes assembly quality less of an issue to answer the question of linear arrangement of ParaHox genes in this species. In order fix these issues and get actual genomic locations of the genes in question, more complete assemblies need to be made, sequencing high molecular weight genomic DNA with high-coverage, employing Next Generation Sequencing (NGS) technologies to increase read size and to account for DNA incontiguity and for low-complexity, repeating regions, and then annotating these assemblies. Such NGS technologies include Pacific Biosciences PacBio sequencing, which allows reads over 10 kb with N50 over 20 kb, thus accounting for repetitive regions (Rhoads & Au 2015), as well as using proximity data such as HiC to help assemble the genome contigs into complete chromosome-sized scaffolds (Dudchenko et al. 2017; Peichel et al. 2017). This will, indeed, give precise genomic locations of the genes and allow for evolutionary comparisons, taking into account similarities and differences of the linear gene organization on genomic scaffolds.

In order to assess three dimensional organization of the chromatin around the loci of interest, chromosome conformation capture (3C) based technologies can be utilized (Dekker et al. 2002; Fraser et al. 2015; Davies et al. 2017). They involve formaldehyde fixation to preserve chromatin organization for the later steps. Chromatin DNA is then cross-linked and cut by restriction enzymes, further, the cut chromatin is ligated so that regions of chromatin that were close to each other in the three dimensions would be ligated together. DNA is then de-cross-linked, extracted and sequenced. The analysis of the sequenced DNA library allows determining what regions of chromatin were close to each other in the nucleus (Davies et al. 2017). Original 3C method can identify interactions between selected loci (Dekker et al. 2002), methods developed based on 3C were designed to increase throughput by identifying more interactions per one library. 4C (circularized conformation capture) is used to identify all the interactions of a selected locus with the rest of the genome (Zhao et al. 2006). Hi-C, on the other hand, is truly high-throughput and can be used to obtain the genome wide interactions of all loci in the genome (Lieberman-Aiden et al. 2009). In *S. purpuratus,* ParaHox genes are dispersed so there are three points of interest to use for the HiC analysis, one for each ParaHox gene: *SpLox*, *SpCdx* and *SpGsx*. No such analysis has previously been done for ParaHox genes, but similar approach was used for Hox genes in amphioxus (Acemel et al. 2016) using 4C data in amphioxus and HiC data in a vertebrate to compare HoxA and HoxD clusters, to suggest that vertebrate clusters exhibit more long-range interactions. Performing similar analysis for the sea urchin species, that lack clustering of the ParaHox genes, will give insight into the organization of ParaHox

genes in the genomes of deuterostomes, highlighting differences between nuclear organization of ParaHox genes in different species and shedding light on the evolution of regulation of these genes.

In addition to the relative three dimensional organization of genes *Lox*, *Cdx* and *Gsx* in the nucleus, the accessibility of DNA around these loci also plays a crucial role in the regulation of their expression, since regulatory DNA is likely to be found in these regions (John et al. 2011). There is a number of techniques to assess chromatin accessibility such as MNase-seq, DNase-seq, Formaldehyde-Assisted Isolation of Regulatory Elements followed by sequencing (FAIRE-seq) and Assay for Transposase-Accessible Chromatin with next-generation sequencing (ATAC-seq) (Tsompana & Buck 2014). MNase-seq allows exploring nucleosome distribution across the genome by allowing to sequence nucleosome bound DNA. DNase-seq, FAIRE-seq and ATAC-seq allow to gain insight into the nucleosome-free regions of the genome. ATAC-seq was developed by Dr Jason Buenrostro and colleagues in 2013 (Buenrostro et al. 2013). Requiring less than 150000 cells, ATAC-seq utilizes a hyperactive Tn5 transposase to cut open chromatin and ligate sequencing adapters at the sites of the cut. The obtained fragments can be amplified via a PCR reaction to produce a sequencing library in under three hours (Buenrostro et al. 2015; Magri et al. n.d.). Sequenced ATAC-seq libraries can be mapped onto a reference genome to infer regions of increased accessibility by identifying regions with more reads mapped (ATAC-seq peaks). These regions are concordant with DNase-seq produced peaks as well, while requiring lower amount of material (Buenrostro et al. 2013), and they give a

genome-wide picture of chromatin accessibility allowing identification of putative cis-regulatory modules.

The following subsections of this chapter will concern the genomic locations of ParaHox genes deduced through the use of newly sequenced and assembled genomes, the three-dimensional organization and interactions of ParaHox genes within the chromatin, as well as assessment of open chromatin regions around these genes to predict putative CRMs in the four species of interest, allowing to gain evolutionary insight into the control of ParaHox genes.

3.2 Results

3.2.1 Chromosomal organization of the ParaHox genes

The first ParaHox cluster was described in 1998 in *Branchiostoma floridae* however there was no information on the organization of ParaHox cluster in *Branchiostoma lanceolatum* prior to the publication of the *B. lanceolatum* genome in 2018 (Marlétaz et al. 2018). This genome was also annotated, so the locations of the genes of interest are easily accessible. The linear organization of the *B. lanceolatum* ParaHox gene cluster is almost identical to that of *B. floridae*: *BlGsx* is the most 5' located gene on the chromosome, followed by *BlLox* (*Lox* gene is also called *Xlox*) 26 kb away and then followed by the 3'-most gene *BlCdx* mere 2.5 kb away from *BlLox* (Marlétaz et al. 2018) (Figure 3.1). This short distance is somewhat in contradiction with the information available for *B. floridae* where the distance between *Lox* and *Cdx* is 7.5 kb (Brooke et al. 1998). The relative orientation of the genes is the same in both species: *Gsx* and *Lox* are transcribed in the same direction while *Cdx* is transcribed in the opposite direction.

Significance of the lower genomic distance between *Lox* and *Cdx* in one species compared to the other is unclear, since both species show the same pattern of spatial and temporal order of expression of these genes. However, decreased distance potentially allows using the same enhancer by the two *Lox* and *Cdx* genes.

In the *Patiria miniata* genome the three genes are also located in a cluster (Annunziata et al. 2013), showing the same organization as in amphioxus. However, the genomic distances between the genes are different: *PmLox* and *PmCdx* are 13 kb apart from each other, while in the genome assembly 2.0 *PmGsx* is located on a different scaffold (these scaffolds are connected by a dashed line in Figure 3.1). *PmLox* and *PmCdx* are on a scaffold that is 42 kb in size, with 40 kb without any gene annotations from the start of the scaffold to the *PmLox* gene. *PmGsx* scaffold also has a region with no gene annotation from *PmGsx* to the end of the scaffold. This region is 70kb long, and, considering that ParaHox genes in this sea star species is arranged in a cluster, it is likelly that these annotation free regions are the region between *PmLox* and *PmGsx*. Thus, there seems to be an issue with the *P. miniata* assembly 2.0 since it was shown via BAC sequencing that there is 31 kb between *PmGsx* and *PmLox* (Annunziata et al. 2013). Transcriptional orientations of the ParaHox genes are the same as in the two amphioxus species discussed. *PmGsx* and *PmLox* share transcriptional orientation, but it is reversed for *PmCdx* (Figure 3.1), highlighting this similarity and indicating that such linear organization and transcriptional directions are ancestral, which has been suggested before (Annunziata et al. 2013).

In case of the sea urchin species, in the genome assembly 3.1, the three genes are on different scaffolds (Sea Urchin Genome Sequencing Consortium et al. 2006), while no *P. lividus* genome was publically available. Dr Arnone (director of studies for this project) has participated in the sequencing consortia to sequence and assemble high-quality genomes for these echinoderm species. *P. lividus* genome, which was sequenced using Illumina, PacBio and HiC technologies, became available in February 2018 is 927.48 megabases (Mb) 3,747 scaffolds, with scaffold N50 being 41.5Mb (Marlétaz. personal communication). *S. purpuratus* genome version 5.0 became available only in late June 2019. The genome was produced using 140x PacBio coverage and 75x HiC coverage resulting in a 845 Mb assembly with an N50 of 37 Mb (Cary. personal communication). The *P. lividus* assembly was annotated by Dr Marlétaz using an Augustus-EVMP-PASA pipeline, using a *P. lividus* transcriptome assembly as well as sea star and amphioxus transcripts and protein data (Marlétaz. personal communication). The *S. purpuratus* genome 5.0 had no gene annotations, so a draft annotation file was made by mapping *S. purpuratus* version 3.1 genes onto the new genome assembly (see section 2.20 and Table 2.3). The newly generated assemblies support the notion that the ParaHox genes in sea urchin are not located in a cluster. In *P. lividus* these genes are over 3 Mb away on a scaffold that is 40 Mb large (Scaffold_3428). *PlGsx* is 5'-most gene which is around 3.4 Mb away from *PlLox* which is the middle gene on the scaffold. *PlLox* in turn is 12.6 Mb away from *PlCdx* which is the 3'-most gene on the scaffold (Figure 3.1). The regions between the ParaHox genes contain other gene annotations. Transcriptional direction, compared to sea star and amphioxus, is

also changed: *PlLox* and *PlCdx* are transcribed in the same direction, while *PlGsx* is transcribed in the opposite direction to the other two ParaHox genes (Figure 3.1). This evidence shows that there is no ParaHox cluster in *P. lividus* and their relative transcriptional orientations are changed, despite the same gene order.

S. purpuratus genome 5.0 annotation shows that, indeed, *S. purpuratus* lacks an intact ParaHox cluster, as was suggested since these genes were identified in sea urchin and the first genome drafts were assembled. In the new genome assembly they are located on a single chromosome-sized scaffold (HiC_scaffold_11, 33.3 Mb in size) (Figure 3.1). However, compared to other species of interest for this thesis, the order of these genes on the scaffold is changed. The 5'-most gene on HiC_scaffold_11 is *SpLox* located 9.9 Mb from the next gene, *SpGsx*, which, in turn, is 6.1 Mb far from the third gene on the scaffold, which is *SpCdx* (Figure 3.1). Transcriptional orientations of these genes are also changed compared to *B. lanceolatum*, *P. miniata* and *P. lividus*, since all three genes are transcribed in the same orientation (Figure 3.1). This suggests that once the cluster is broken, keeping the same transcriptional annotation is not evolutionary necessary. ParaHox genes in the two sea urchin species are expressed in the same way developmental stage-wise and in the same locations. It was suggested that presence of an intact ParaHox cluster is unnecessary for spatial linearity of the ParaHox gene expression (*Gsx*- anterior, *Lox*- middle, *Cdx*- posterior), but may play a role in the order of expression onset during development (Ikuta et al. 2013; Annunziata et al. 2013).The data presented in this section, however, also suggests that gene locations on the scaffold and their relative transcriptional orientations are also not important for spatial linearity (in

fact, using collinearity for sea urchin species seems wrong as gene locations are not collinear with expression) but still may play a role in maintaining the temporal collinearity.

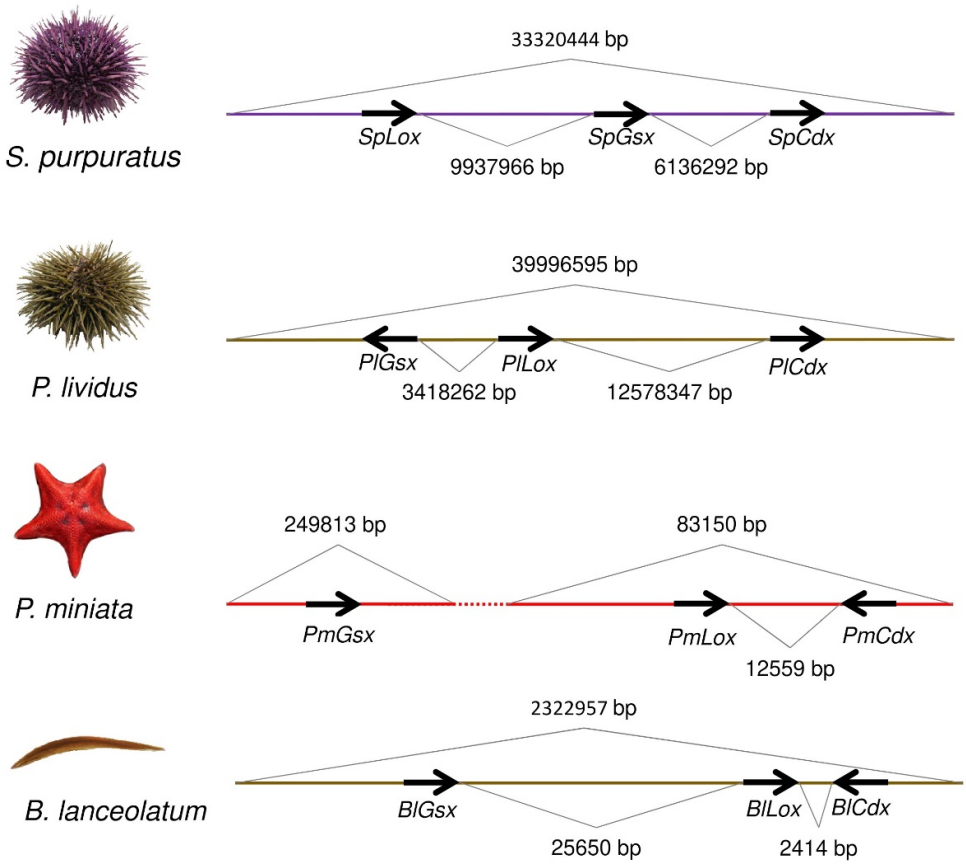


Figure 3.1 Chromosomal organization of the ParaHox genes in the four species of interest. Scaffold sizes, distances between the genes and their relative orientations based on genome assemblies are shown. Dashed line for *P. miniata* indicates that the two scaffolds were shown to be contiguous.

3.2.2 Three-dimensional organization of the ParaHox loci in S. purpuratus

The ParaHox genes in the sea urchin species are dispersed and are located at chromosomal loci that are far away from each other. However, due to chromatin packing and looping the three genes might still be close to each other, and might form a co-regulated gene group. In order to examine the three dimensional organization of chromatin around the ParaHox loci, the HiC approach was used (see section 2.17 and 3.1). This approach allowed to identify intra- and interchromosomal interactions within the *S. purpuratus* genome at 100 kb resolution, and, in particular, interactions between the ParaHox loci at 20 kb resolution (Table 3.1 and Figure 3.2).

Table 3.1 HiC datasets information used to assess *S. purpuratus* chromatin three-dimensional structure.

S. purpuratus				
HiC				
Sample	Number of reads	Alignment rate	Number of valid interactions ParaHox (20 kb)	Number of valid interactions Whole Genome (100 kb)
50hpf forward reads	648710836	57.50%	5542	27444341
50hpf reverse reads	648710836	55.46%		

The *S. purpuratus* ParaHox genes are located in separate topologically associated domains (TADs) with relatively few inter-domain interactions (Figure 3.2 A and C). The HiC contact map for each of the genes of the ParaHox genes (Figure 3.2 A and C) show that there are more physical interactions between the ParaHox genes and the loci proximal to them on the scaffold, than with more

distal regions on the same scaffold. The analysis also shows the divisions of TADs into sub-TADs in which the contacts are even more frequent (Figure 3.2 A and C).

The contact matrices and maps show that there are no connections between *SpLox* and *SpCdx* loci, however both of these loci do show some interaction with *SpGsx* locus (Figure 3.2 B). However, at 20 kb resolutions these interactions are very sparse: connection between *SpLox* and *SpGsx* resulted in five counts in the contact matrix, while *SpCdx* interaction with *SpGsx* gave only one count (indicated by different contact line width in Figure 3.2 B). The sparsity of these counts points to these interactions being insignificant in the context of the whole genome. Unfortunately, there are few tools available to test the significance of the interactions of binned genomic locations to the author's knowledge, such as HiC-DC and GOTHiC (Carty et al. 2017; Mifsud et al. 2017). However, these tools require a BSgenome R object in their pipelines, which is not currently available for *S. purpuratus*. Another tool available for identifying significant interactions from HiC data, HOMER (Heinz et al. 2010), highlighted that no identified interaction of the ParaHox loci were significant (see Non-book component). Therefore, with caution, it is possible to suggest that there are no significant interactions between the ParaHox genes, and that the obtained counts could be obtained randomly. In addition, since no interactions were identified between the two digestive system genes *SpLox* and *SpCdx*, there is likely to be no contact between their genomic loci in the chromatin three-dimensional arrangement. Thus, they are unlikely to be co-regulated via physical proximity. Therefore, cis-regulatory modules regulating these genes need to be considered separately,

and thus *Gsx* location is unlikely to have an impact on regulation of the gut related genes. Therefore, the following sections and chapters, will focus solely on the gut genes: *Lox* and *Cdx*.

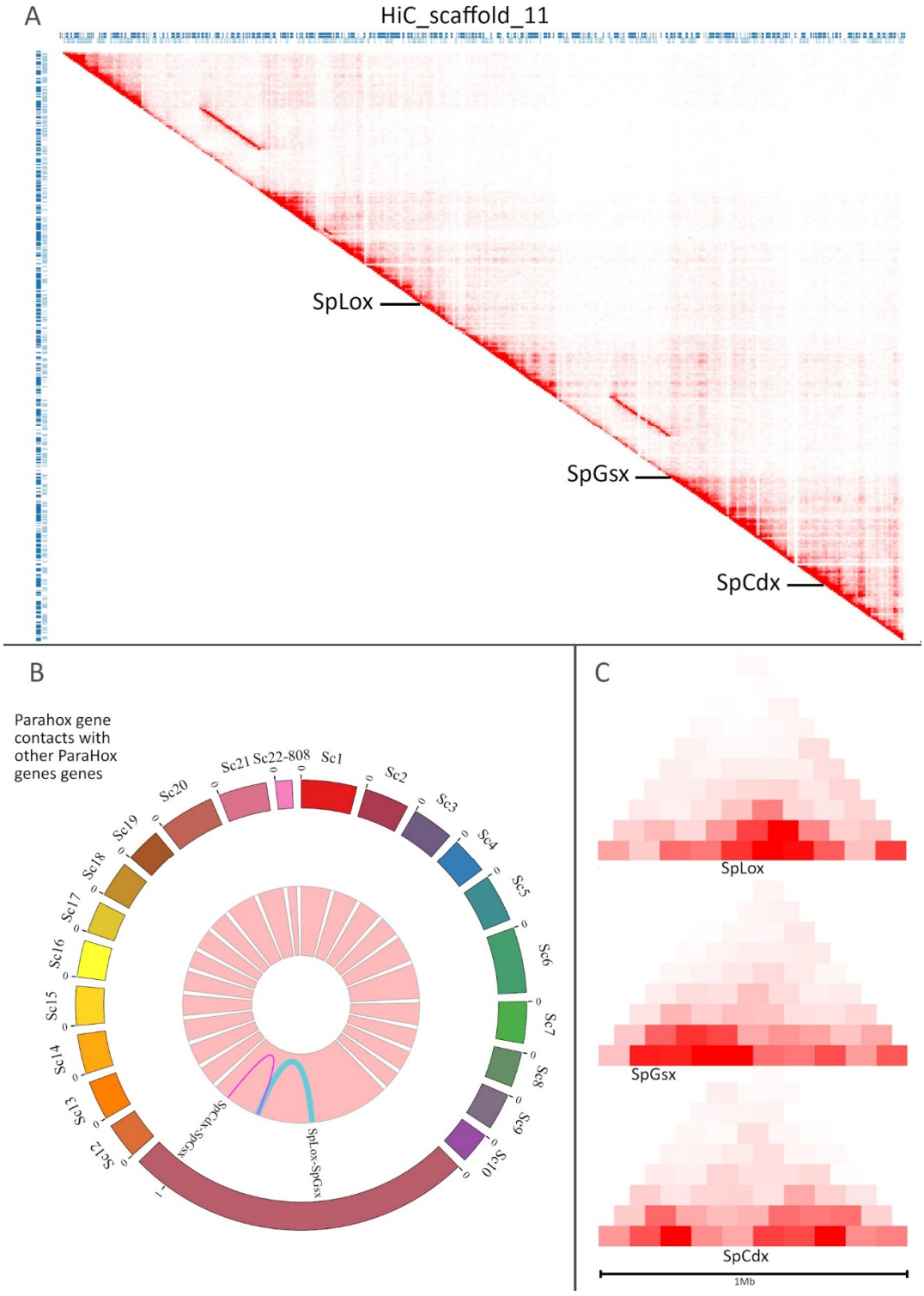


Figure 3.2 *S. purpuratus* ParaHox HiC results. A. HiC contact map of the scaffold, on which the ParaHox genes are located. Loci of the ParaHox genes are marked. Note the red lines parallel to the diagonal, these lines are likely due to assembly errors. B. BioCircos circular map of the ParaHox loci interactions between each other. The relative thickness of the connecting lines

represents the relative number of counts in the contact matrix. HiC_scaffold_11 (Sc11) is represented as 10x larger than other scaffolds to allow visualization. C. Parts of topologically associated domains containing ParaHox loci at 100 kb resolution. Each region shown is 1 Mb divided into smaller sub-domains, locations of ParaHox genes, *SpGsx*, *SpLox* and *SpCdx*, are labeled by gene names.

3.2.3 ATAC-seq identifies regions of open chromatin

In order to assess chromatin accessibility around the ParaHox genes, we have generated genome-wide assays of open chromatin using ATAC-seq for the three species of echinoderms. In the purple sea urchin *S. purpuratus,* the late blastula stage corresponds to 24 hours post fertilization (hpf), when cultured at 15°C, late gastrula is at 48 hpf, while prism and pluteus are at 66 hpf and 72 hpf, respectively. For *P. lividus* that were collected in the Mediterranean Sea and cultured at 18°C the blastula is at 15 hpf, gastrula at 24 hpf, while pluteus is at 40 hpf. In case of the sea star *P. miniata,* comparable stages of blastula, gastrula and bipinnaria larva are at 24, 66 and 90 hpf, respectively. Genome wide ATAC-seq assays were performed for these time points in two biological replicates. In addition, for *S. purpuratus,* gut samples at 48 hpf and 66 hpf were extracted and used to create gut tissue specific ATAC-seq libraries (For embryo cultures, tissue extraction, library preparations, reads mapping and replicate treatment see sections 2.1-2.3 and 2.11) (Table 3.2).

Table 3.2 ATAC-seq datasets information for the four species of interest. Peaks that overlap in both samples of a particular timepoint represent the concordant set.

S. purpuratus				
ATAC-seq				
Sample	Number of reads	Alignment rate	Number of peaks	Number of concordant peaks

24hpf whole embryo A	40470592	51.13%	46036	24907
24hpf whole embryo B	40678575	45.36%	39037	
48hpf whole embryo A	33751915	51.98%	33860	23979
48hpf whole embryo B	34752400	50.25%	38602	
48hpf gut tissue A	85229316	41.84%	47131	31850
48hpf gut tissue B	73807090	55.22%	72952	
66hpf whole embryo A	67270304	46.38%	43260	33749
66hpf whole embryo B	67737883	48.75%	65144	
66hpf gut tissue A	47769078	50.52%	51847	3963
66hpf gut tissue B	47156190	36.80%	6911	
72hpf whole embryo A	62971524	36.05%	45130	19068
72hpf whole embryo B	78821537	33.43%	26577	
P. lividus				
ATAC-seq				
Sample	Number of reads	Alignment rate	Number of peaks	Number of concordant peaks
15hpf whole embryo A	48142078	51.90%	85660	60411
15hpf whole embryo B	48791083	51.11%	90595	
24hpf whole embryo A	33858822	49.36%	67624	46893
24hpf whole embryo A	41251274	51.04%	75735	
40hpf whole embryo A	62030520	50.85%	101676	77091
40hpf whole embryo A	72684938	51.60%	115687	
P. miniata				
ATAC-seq				
Sample	Number of reads	Alignment rate	Number of peaks	Number of concordant peaks
24hpf whole embryo A	34864472	68.70%	87749	66638
24hpf whole embryo B	33490150	69.15%	85310	
66hpf whole embryo A	30696732	63.74%	85010	61403
66hpf whole embryo B	24953519	66.06%	78761	
90hpf whole embryo A	45053710	60.47%	85736	70036
90hpf whole embryo B	57810590	49.50%	104989	
B. lanceolatum				
ATAC-seq				
Sample	Number of reads	Alignment rate	Number of peaks	Number of concordant peaks
8hpf whole embryo A	Concordant peak sets obtained from Dr Gómez Skarmeta			25455
8hpf whole embryo B				
15hpf whole embryo A				47002
15hpf whole embryo B				
36hpf whole embryo A				39537
36hpf whole embryo B				

This approach allowed to identify open chromatin regions which give “peaks” after mapping onto a genome assembly, since many reads would map at that region (Table 3.2 and Figure 3.3).

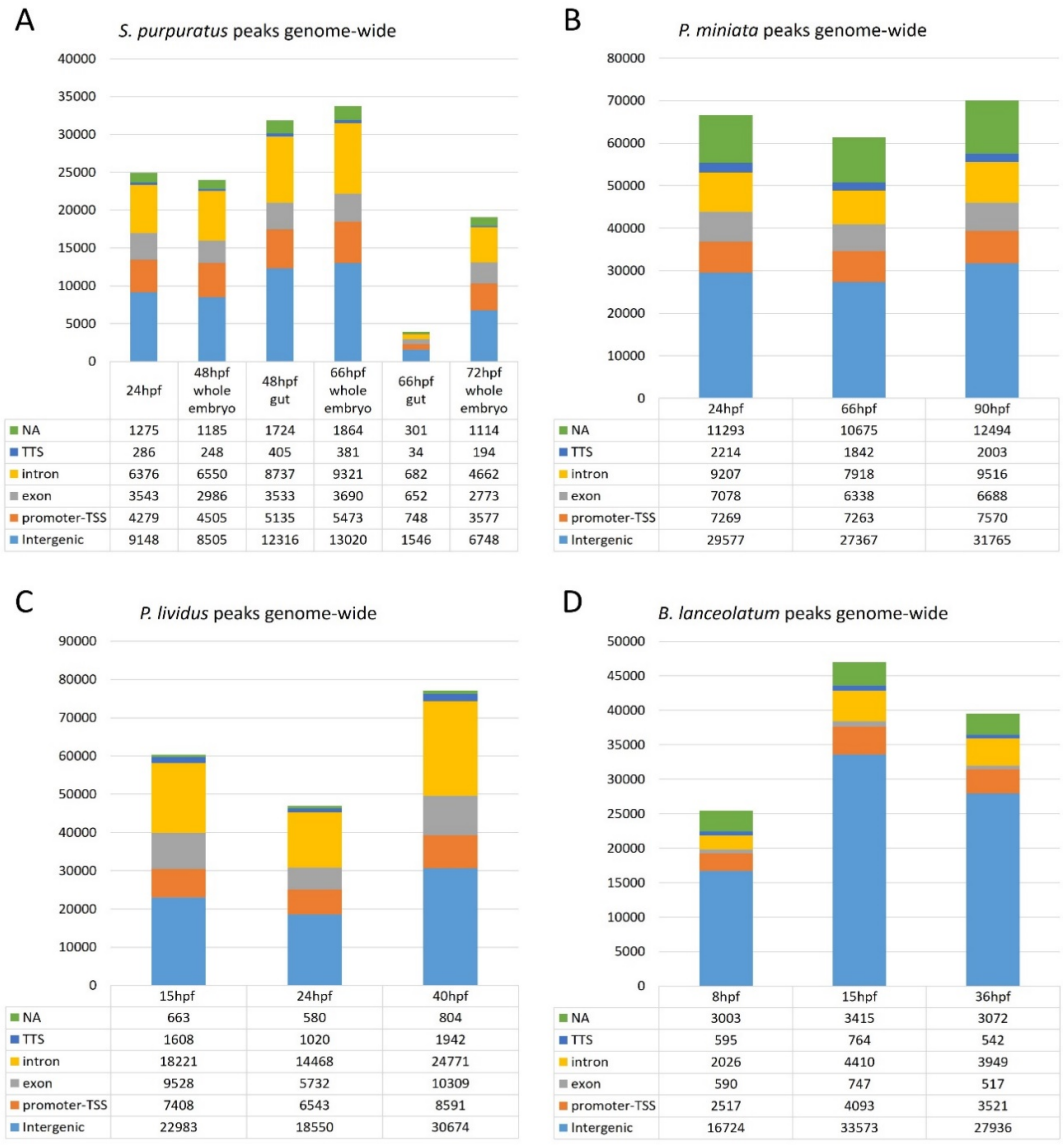


Figure 3.3 ATAC-seq peak distribution in relation to gene annotations. A. Distribution of ATAC-seq peaks in the genome of *S. purpuratus* version 3.1 relative to intergenic regions, promoters/transcription start sites, gene exons, introns and transcription termination sites, as well as peaks that could not be attributed to any of these features. B. Distribution of ATAC-seq peaks in the genome of *P. miniata* version 2.0 relative to intergenic regions, promoters/transcription start sites, gene exons, introns and transcription termination sites, as well as peaks that could not be

attributed to any of these features. C. Distribution of ATAC-seq peaks in the genome of *P. lividus* relative to intergenic regions, promoters/transcription start sites, gene exons, introns and transcription termination sites, as well as peaks that could not be attributed to any of these features. D. Distribution of ATAC-seq peaks in the genome of *B. lanceolatum* relative to intergenic regions, promoters/transcription start sites, gene exons, introns and transcription termination sites, as well as peaks that could not be attributed to any of these features. Intergenic regions are light blue, promoter-TSS regions are brown, exons are grey, introns are orange, transcription termination sites (TTS) are dark blue and unattributable regions (NA) are green.

For *S. purpuratus,* the ATAC-seq assay showed 24907 open chromatin regions at the blastula stage. At the gastrula stage, 23979 open chromatin regions were identified in total, with gut specific ATAC-seq at the same developmental stage resulting in 31850 open chromatin regions. At later stages of development of the purple sea urchin larva, the ATAC-seq assay provided 33749 regions of open chromatin at the prism stage, while the gut specific samples only gave 3963 loci, where DNA is accessible, after replicate combination. At the pluteus stage, the assay identified 19068 peaks corresponding to accessible chromatin (Table 3.2). These loci are mostly found in the intergenic regions and introns of genes (Figure 3.3 A). However, for some of these regions it was impossible to identify where they are in relation to gene annotations, due to mapping onto scaffolds without any genes identified in the genome assembly. The *S. purpuratus* genome 3.1 was used for analysis of the ATAC-seq and other libraries due to the availability of high-quality gene models, transcriptome and proteome compatibilities, and, therefore, to ensure consistency across different libraries generated such as ATAC-seq, RNA-seq or scRNA-seq. However, as mentioned before the 3.1 genome assembly is incomplete with multiple gaps, which results in low mapping rates (Table 3.2), so a fraction of ATAC-seq peaks ended up mapped to smaller gene-less scaffolds (NA in Figure 3.3).

Open chromatin regions for a Mediterranean sea urchin *P. lividus* were also identified. At the blastula stage, the resulting total number of peaks was 60411, while the gastrula assay resulted in 46893 peaks and the pluteus in 77091 (Table 3.2). The peaks are also mostly distributed in regions between genes and within introns (Figure 3.3 C) as in *S. purpuratus*. It is worth noting the relatively low number of open chromatin regions in scaffolds where the region's relative location to a gene is impossible to identify (NA in Figure 3.3 C), highlighting a better quality of assembly compared to the *S. purpuratus* genome assembly version 3.1.

Regions of accessible chromatin in *P. miniata* follow the same pattern with most of the peaks located in the intergenic stretches on the genomic scaffolds and in the introns (Figure 3.3 B). At the blastula stage, 66638 open chromatin regions were identified, while the sequenced libraries for the gastrula resulted in 61403 accessible DNA loci and 70036 open chromatin loci were also identified from the bipinnaria larva ATAC-seq (Table 3.2 and Figure 3.3 B). While the mapping rates are relatively high (Table 3.2), it is worth noting that the proportion of peaks in the scaffolds without genes is also high compared to other echinoderm species (Figure 3.3 B). This, unlike in *P. lividus*, could indicate an assembly of worse quality than *S. purpuratus* genome 3.1, which was already discussed in relation to the ParaHox cluster being broken into two different scaffolds in the assembly despite the short distances (less than 30 kb) between the genes of the cluster (see section 2.1).

Open chromatin data, generated via ATAC-seq method, for the amphioxus *B. lancelatum* was obtained from Dr Gómez Skarmeta. The data for embryonic stages of *B. lanceolatum* comparable to embryonic stages used to generate echinoderm ATAC-seq libraries was requested. In *B. lanceolatum,* 8 hours post fertilization corresponds to the blastula stage, 15 hpf is a late gastrula or early neurula, while 36hpf is an early larva of the amphioxus. At the blastula stage, ATAC-seq experiments identified 25455 open chromatin regions, with 47002 regions at the late gastrula stage, and 39537 loci of accessible DNA at the early larva stage (Table 3.2). Majority of these regions are between genes, and, again, it was impossible to identify any gene-relative information for some of the peaks. For the current *B. lanceolatum* assembly the issue of short gene annotation free scaffolds also still exists. It is worth noting that the percentage of peaks identified in the introns is much smaller compared to other species examined, this could be due to shorter length of introns in the amphioxus genome, since the relative percentage of intergenic regions in all species in question is similar, around 30% of the total assembly, or could be a feature of amphioxus chromatin. In terms of the relative number of peaks in the introns, *P. miniata* is somewhat between the sea urchins and amphioxus, this could reflect the differences in gene structures (Figure 3.3).

Number of peaks is hard to compare between species since there are a lot of factors that affect it, such as nuclei quantity, library quality, sequencing depth, polymorphisms and assembly completeness. The number of nuclei is important for the enzymatic transposition to generate the ATAC-seq libraries, if the number of nuclei is too low, then the over-tagmentation of material is possible, which

would lead to high background during read mapping, which would cause issues in peak calling. It is possible that in some cases, such as the *S. purpuratus* 66hpf gut samples, the number of nuclei was not sufficient to generate a high-quality ATAC-seq library. Insufficient sequencing depth can also result in high background as it may not be representative of the library. Polymorphisms make mapping difficult due to sequence differences between the library and the reference genome, while taxa of interest are known for high polymorphism (Putnam et al. 2008; Marlétaz et al. 2018; Sea Urchin Genome Sequencing Consortium et al. 2006), thus low mapping rates of the sea urchin ATAC-seq datasets could be attributed to polymorphism (Table 3.2). Due to incomplete assemblies, some of the reads will not be able to map anywhere on the assembly if they fall into the gap regions, thus not contributing to the total peak count.

Relative locations of the peaks compared to the genes are also of functional importance. Open chromatin regions are associated with cis-regulatory modules (John et al. 2011), which can be located upstream of the promoter, downstream of it, and in the gene bodies. The distribution of the peaks obtained during this thesis were also found to be significant and not random (p-value= 0. for each species and timepoint) via the chi-squared test described in subsection 2.22.1, this suggests that the distribution of identified open chromatin regions has a biological meaning. In the sea urchin, the relative abundance of peaks in the introns highlights their potential of being associated with CRMs, since these regions are cumulatively shorter than intergenic regions. Open chromatin at promoter regions have been associated with expressed genes (Wang et al. 2012) thus it is expected that ATAC-seq peaks would map to promoter-transcription

start site (TSS) regions, which is supported by the generated ATAC-seq datasets (Figure 3.3) and the chi-squared test performed on these datasets (p-value= 0. for each species and every timepoint for promoter-TSS). In addition, there is a known negative correlation between the active genes and the open chromatin at the 3' ends of the genes (Wang et al. 2012), also supported by the low fraction of peaks falling near transcription termination sites (TTS) (Figure 3.3) A fraction of open chromatin regions are also found within exons and the existence of exonic enhancers (Li et al. 2015; Birnbaum et al. 2012; Ritter et al. 2012; Neznanov et al. 1997) shows that coding regions may also be associated with cis-regulatory modules and thus this fraction of the peaks of the ATAC-seq data could also be functionally important.

During the development, a large number of genes is employed and this employments needs to be tightly regulated to give rise to different tissue and cell types, through dynamic processes with genes being "turned on" and "turned off". Throughout the developmental stages used to obtain the ATAC-seq data, majority of the genes annotated have ATAC-seq peaks attributable to them (Figure 3.4). In *P. lividus* and *P. miniata* this number is close to 80% (Fig 2.4 B and C), while in *S. purpuratus* it is 67% (Figure 3.4 A), which corresponds to approximate percentage of genes expressed at different stages of development which is around 72% of all genes (Tu et al. 2012). However, the significance tests indicated that the number of genes with peaks is unlikely to have much biological meaning, since the p-value for the significance testing in *S. purpuratus* is 0.96502551 and 0.22073841 in *P. lividus*, despite the fact that the chi-squared tests showed significant differences between genes near actual peaks and genes

near random peaks in *P. minitata* (p-value= 7.35221324e-191) or *B. lanceolatum* (p-value= 3.05434303e-43). The biological meaning of this difference is unclear, but could be related to the relatively worse assembly qualities, if smaller scaffolds affect random peak re-shuffling.

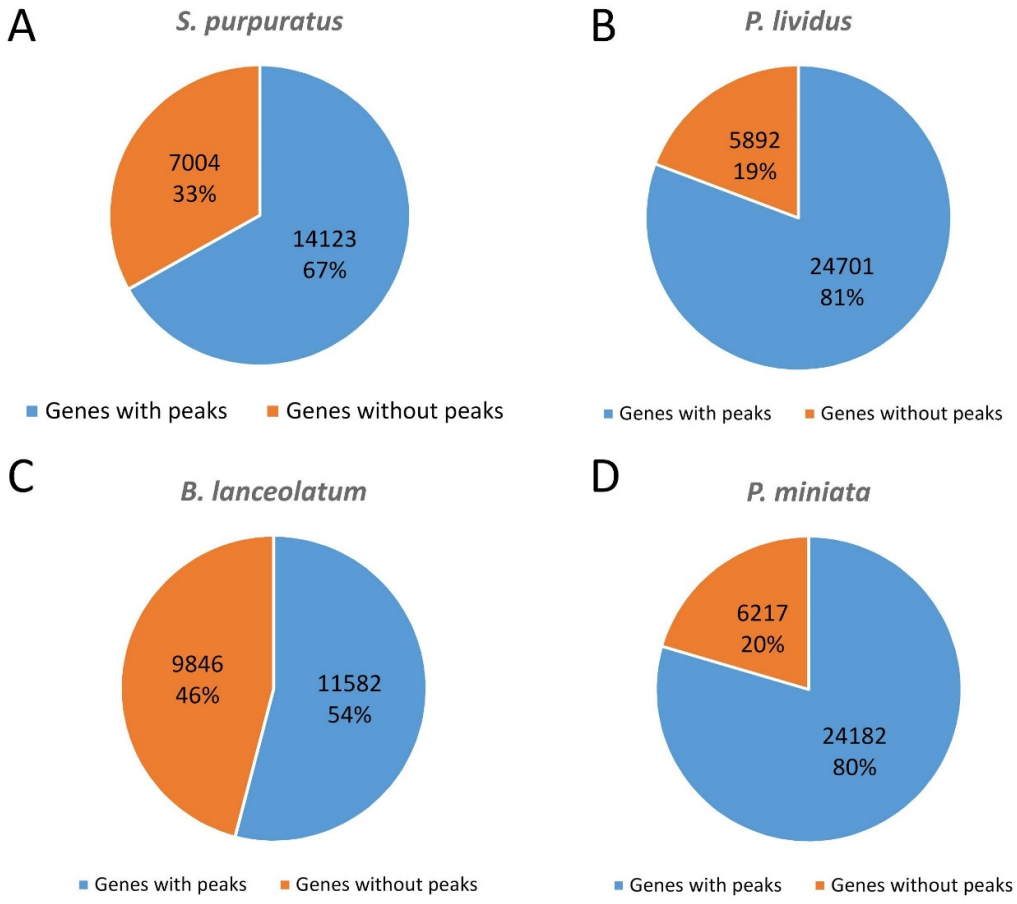


Figure 3.4 Genes with ATAC-seq peaks. A. Number of genes with attributable ATAC-seq peaks also shown as percentage of all genes for *S. purpuratus* datasets. B. Number of genes with attributable ATAC-seq peaks also shown as percentage of all genes for *P. lividus* datasets. C. Number of genes with attributable ATAC-seq peaks also shown as percentage of all genes for *B. lanceolatum* datasets D. Number of genes with attributable ATAC-seq peaks also shown as percentage of all genes for *P. miniata* datasets

Therefore, ATAC-seq allowed generating open chromatin data for the species of interest as well as gain insight into the distribution of the potential cis-regulatory regions in the genomes. The ATAC-seq peak distribution trends are generally consistent across species, which is unsurprising considering that gene regulation through chromatin access is performed through same mechanisms across metazoans, and generated data supports that. Significant percentage of peaks in the promoter-TSS regions points to the high quality of the data and its potential usefulness for CRM prediction.

3.2.4 Known CRMs highlight predictive power of the ATAC-seq data

In order to assess the potential of the generated ATAC-seq datasets to predict putative CRMs, a number of known cis-regulatory modules were examined in relation to the open chromatin regions identified through ATAC-seq. As stated before, *S. purpuratus* is an extensively studied experimental organism due to the ease with which gametes can be obtained to set up synchronous embryonic cultures, the transparency of said embryos and the possibility of doing functional studies, such as assessing activity of cis-regulatory elements using reporter constructs and gene knock-down experiments using morpholinos. Literature search allowed to select 43 different cis-regulatory modules, which were shown to contribute to gene expression regulation of 18 genes that play a role in the development of different *S. purpuratus* tissues, cell types and organs. The genes are *SpHox11/13B, SpCyIIa, SpCyIIIa, SpEndo16, SpDelta, SpGatae, SpTbrain, SpSM50, SpSM30α, SpAlx1, SpOtx, SpPks1, SpGcm, SpFoxA, SpNodal, SpCycD, SpWnt8, SpBlimp1* (Table 3.3).

These genes play roles in defining a wide selection of *S. purpuratus* tissues, since *SpHox11/13B* (Cui et al. 2017a), *SpEndo16* (Yuh et al. 2001) and *SpFoxA* (de-Leon & Davidson 2010) all play a role in the development of the embryonic gut. Similarly to *SpOtx* (Yuh et al. 2004), *SpBlimp1* (Livi & Davidson 2006; Smith et al. 2008) and *SpGatae* (Lee et al. 2007), which are early regulators of endoderm specification, showing expression in the endodermal precursors from the early blastula stage and then in the archenteron (Lee et al. 2007) from the gastrula to the pluteus. In case of *SpCycD,* expression is not only in the gut but also in the oral ectoderm and the ciliary band. *SpWnt8* controls the endomesoderm formation: both the gut and the skeletogenic mesoderm (Wikramanayake et al. 2004; Minokawa et al. 2005). *SpTbrain* (Wahl et al. 2009), *SpSM50* (Makabe et al. 1995), *SpSM30α* (Akasaka et al. 1994) and *SpAlx1* (Damle & Davidson 2011) are well known to be nodes of the skeletogenic mesenchyme gene regulatory network. *SpGcm* (Ransick & Davidson 2006) is employed in the development of the secondary mesenchyme and its derivatives from the mid-blastula, *SpPks1* is a marker gene for the pigments cells of the sea urchin embryo that start appearing at the gastrula stage and perform immune functions (Calestani & Rogers 2010), while *SpNodal* is an ectodermal gene, defining oral ectoderm from blastula stage onwards (Nam et al. 2007). *SpCyIIa* is a cytoskeletal gene, since it is an actin, however, while it is expressed in the PMC cells at the blastula and the gastrula, its expression shifts to the secondary mesenchyme coelomic pouch cells by the prism stage, and then gets restricted to the midgut and hindgut at the pluteus stage (Arnone et al. 1998). SpCyIIIa, which is another actin gene, controls development of the aboral ectoderm (Coffman et al. 1997), while *SpDelta* starts

as a PMC cells gene but by the late blastula switches its domain of expression to the secondary mesenchyme (Revilla-i-Domingo et al. 2004).

26 of the 43 selected known CRMs (over 60%) either contain or overlap with ATAC-seq peaks of different developmental stages (Figure 3.5 and Table 3.3). This number was found to be significant with p-value of 1.3424457203547e-18 (see section 2.22.2) compared to a random peak set. Of those, 80.8% have peaks from more than one timepoint, with 57.1 % of those exhibiting open chromatin regions at all the developmental stages for which the ATAC-seq data was generated, except the 66hpf gut ATAC-seq, for which the number of peaks is low due to bad replicates. Still, even with the bad replicates one of the CRMs from the list regulating the expression of *SpBlimp1* shows a peak from the 66hpf gut ATAC-seq data (Table 3.3 and Figure 3.6). The function of *SpBlimp1* in the development of the digestive system is known and this gene is an important node of the known gene regulatory network upstream of the ParaHox genes (Annunziata & Arnone 2014). In majority of cases, chromatin around CRMs stays open throughout the developmental stages assessed, which is not altogether unexpected since majority of these genes are expressed throughout the sea urchin development (Tu et al. 2014). On the other hand, CRMs, that showed open chromatin only in the gut-enriched 48hpf dataset highlight the importance of tissue specific data, since the genes that they regulate are endoderm and gut specific, such as *SpEndo16* (Figure 3.6) (Yuh et al. 2001), *SpGatae* (Lee et al. 2007), *SpCycD* (McCarty & Coffman 2013) and *SpHox11/13B* (Figure 3.6) (Annunziata & Arnone 2014; Cui et al. 2017a). It is also worth mentioning that despite the fact that some known CRMs do not overlap with open chromatin

regions (Wahl et al. 2009), it does not mean that their corresponding genes completely lack ATAC-seq peaks attributable to them by visualization or HOMER software. Those peaks might still be active CRMs that need to be validated.

Majority of the CRMs in question were bioinformatically predicted by the authors of the corresponding studies by identifying conserved sequences around the genes of interest with another sea urchin species, *Lytechinus variegatus* (Yuh et al. 2002), using BAC sequences and, later, the genome assembly, the first draft of which became available in 2011 (GenBank accession: GCA_000239495, (Kudtarkar & Cameron 2017)). The most recent paper from 2018 by Shashikant and colleagues describes the use of ATAC-seq assay to gain insight into cis-regulation of the skeletogenic mesenchyme associated genes such as *SpAlx1* (Shashikant et al. 2018), however our data does not show the same peaks as them around *SpAlx1*, since only one of the four regions described for this gene was found in our data at any stage (Table 3.3). This could be due to the fact that they obtained ATAC-seq data from isolated PMC cells (Shashikant et al. 2018), so reads mapping around the skeletogenic genes would be enriched, while our data is either whole embryo or PMC-depleted gut samples. The high percentage of known cis-regulatory modules found to contain or overlap ATAC-seq peaks highlights the predictive potential of the ATAC-seq data in relation to identification of putative CRMs. In addition, the ATAC-seq method does not require sequence comparisons between species, although such comparisons remain useful, and thus ATAC-seq may help identify novel cis-regulatory elements, which may not be conserved among species.

These findings suggest that ATAC-seq experiments have strong cis-regulatory region predictive power. Peaks obtained through such experiments can be described as putative CRMs that can be further analyzed and tested.

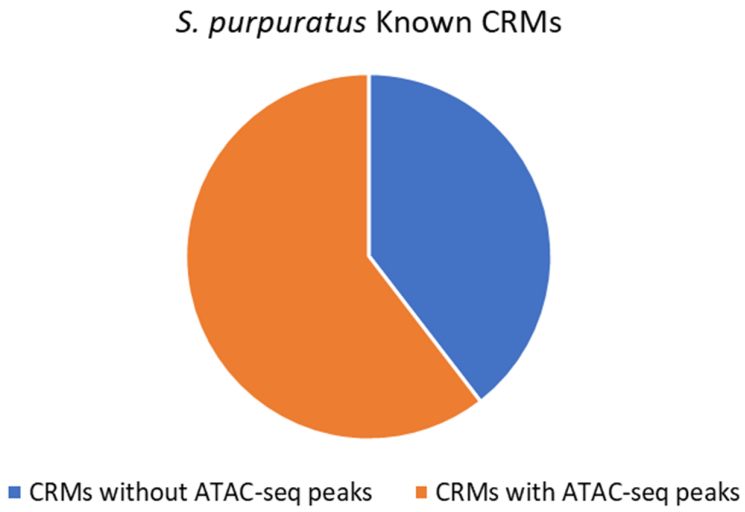


Figure 3.5 Pie-chart showing the proportion of known CRMs with an open chromatin region identified by ATAC-seq and without.

Table 3.3 Known CRMs tested for having an open chromatin region identified by ATAC-seq data.

Scaffold	Start	End	Known CRM	Reference	Presence in ATAC peak
Scaffold636	404163	404759	ME_for_SpHox11/13B	Cui et al. 2017	24hpf, 48hpf, 48hpf_gut, 66hpf
Scaffold636	405533	406206	L_for_SpHox11/13B	Cui et al. 2017	None
Scaffold636	400544	405532	E_for_SpHox11/13B	Cui et al. 2017	48hpf_gut
Scaffold636	459346	462267	D_for_SpHox11/13B	Cui et al. 2017	None
Scaffold1400	163592	165057	AB_for_SpCyIIa	Arnone, et al. 1998	48hpf, 48hpf_gut, 66hpf
Scaffold1400	160584	163597	CDE_for_SpCyIIa	Arnone et al. 1998	None

Scaffold213	671569	674051	Middle_CRM_for _SpCyIIIa	Coffman et al. 1997	24hpf, 48hpf, 48hpf_gut
Scaffold32	569597	571604	B_for_SpEndo16	Yuh et al. 1996 Yuh et al. 2001	None
Scaffold32	571605	571897	A_for_SpEndo16	Yuh et al. 1996 Yuh et al. 2001	48hpf_gut
Scaffold329	976203	979486	R11_for_SpDelta	Revilla-i-Domingo et al. 2004	None
Scaffold127	444647	445321	10_for_SpGatae	Lee et al. 2007	48hpf_gut
Scaffold127	429868	430451	24_for_SpGatae	Lee et al. 2007	48hpf_gut
Scaffold424	292900	293242	gamma_for_ SpTbrain	Wahl et al. 2009	None
Scaffold424	296477	297066	B_for_SpTbrain	Wahl et al. 2009	None
Scaffold424	297995	299594	C_for_SpTbrain	Wahl et al. 2009	None
Scaffold903	122527	123087	CRM_region_for_ SpSM50	Makabe et al. 1995	None
Scaffold4453	25909	28642	CRM_region_for_ SpSM30α	Akasaka et al. 1994	None
Scaffold881	95946	96898	I_for_SpAlx1	Damle and Davidson 2011	None
Scaffold881	94700	95200	CRM1_for_SpAlx 1	Shashikant et al. 2018	24hpf, 48hpf, 48hpf_gut, 66hpf, 72hpf
Scaffold881	91500	92500	CRM2_for_SpAlx 1	Shashikant et al. 2018	None
Scaffold881	90100	91300	CRM3_for_SpAlx 1	Shashikant et al. 2018	None
Scaffold468	439564	442066	Otx15_for_SpOtx	Yuh et al. 2004	24hpf, 48hpf, 48hpf_gut, 66hpf, 72hpf
Scaffold174	179736	182059	CRM_region_for_ SpPks1	Calestani and Rogers 2010	None

Scaffold118	122624	136000	CRM_region_ PED_for_SpGcm	Ransick and Davidson 2006	24hpf, 48hpf
Scaffold345	367073	368013	J_for_SpFoxA	de-Leon and Davidson 2010	24hpf, 48hpf, 48hpf_gut, 66hpf, 72hpf
Scaffold345	368436	368908	I_for_SpFoxA	de-Leon and Davidson 2010	24hpf, 48hpf, 48hpf_gut, 66hpf, 72hpf
Scaffold345	376562	377853	F_for_SpFoxA	de-Leon and Davidson 2010	48hpf_gut, 66hpf
Scaffold345	360253	363946	K_for_SpFoxA	de-Leon and Davidson 2010	24hpf, 48hpf, 48hpf_gut, 66hpf, 72hpf
Scaffold1203	37572	38076	5P_for_SpNodal	Nam et al. 2007	24hpf, 48hpf, 48hpf_gut, 66hpf, 72hpf
Scaffold1203	35431	36175	INT_for_SpNodal	Nam et al. 2007	24hpf, 48hpf, 48hpf_gut, 66hpf, 72hpf
Scaffold1203	33766	34428	3P_for_SpNodal	Nam et al. 2007	None
Scaffold87	253204	257105	2_for_SpCycD	McCarty and Coffman 2013	48hpf, 48hpf_gut, 66hpf
Scaffold87	250003	250720	4-1_for_SpCycD	McCarty and Coffman 2013	24hpf, 48hpf, 66hpf, 72hpf
Scaffold87	248737	249175	4-2_for_SpCycD	McCarty and Coffman 2013	24hpf, 48hpf, 48hpf_gut, 66hpf, 72hpf
Scaffold87	243659	246065	5_for_SpCycD	McCarty and Coffman 2013	48hpf_gut
Scaffold87	240263	242947	6_for_SpCycD	McCarty and Coffman 2013	None
Scaffold87	246278	248351	17_for_SpCycD	McCarty and Coffman 2013	24hpf, 48hpf, 48hpf_gut, 66hpf, 72hpf
Scaffold87	229741	234383	19_for_SpCycD	McCarty and Coffman 2013	48hpf, 48hpf_gut
Scaffold102	167151	167304	C_for_SpWnt8	Robertson et al. 2008	None
Scaffold102	175708	176485	A_for_SpWnt8	Minokawa et al. 2005	24hpf, 48hpf, 48hpf_gut, 66hpf

Scaffold1008	99633	101557	43_for_SpBlimp1	Livi and Davidson 2007	24hpf, 48hpf, 48hpf_gut, 66hpf, 72hpf
Scaffold1008	87232	88109	CR2_for_SpBlimp1	Smith et al. 2008	24hpf, 48hpf, 48hpf_gut, 66hpf, 66hpf_gut, 72hpf
Scaffold1008	79077	79943	CR5_for_SpBlimp1	Smith et al. 2008	24hpf, 48hpf, 48hpf_gut, 66hpf, 72hpf

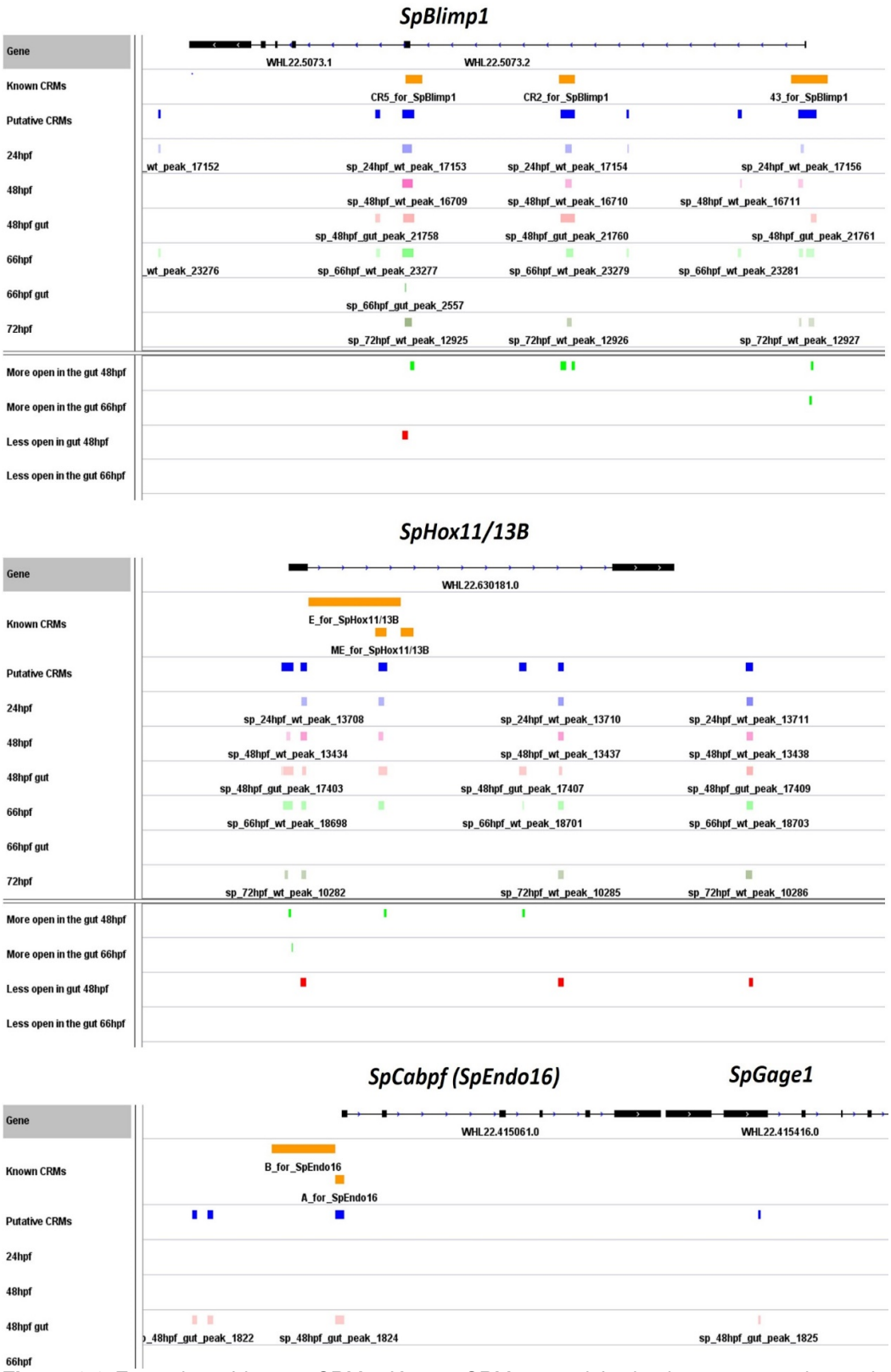


Figure 3.6 Examples of known CRMs. Known CRMs tested for having an open chromatin region identified by ATAC-seq data. Gene annotations, known and putative CRM locations and ATAC-seq peaks shown.

3.2.5 Differential ATAC-seq analysis reveals regions more open in the gut

Gut enriched ATAC-seq data sets were obtained by separating sea urchin gut tissue by glycine-EDTA treatment followed by mechanical separation by pipetting in calcium-magnesium-free artificial sea water (see section 2.2), and using these samples for the ATAC-seq library preparation (see section 2.3). Therefore, these samples differ from the whole embryo sample in that they should lack the ectoderm, the blastocoeliar cells and the majority of the primary mesenchyme cells. Due to the fact that this data set is merely gut enriched compared to the whole embryo (since the whole embryo contains all cells and tissues, including the gut), the whole embryo ATAC-seq would contain most of the peaks present in the gut data since peak calling is performed independently, which points to the need of differential analysis of the gut ATAC-seq samples compared to the whole embryo datasets. In order to do this, overlapping peaks from the two conditions (gut and whole embryo) at the same timepoint were merged and broken up into 150 bp windows (roughly equalling nucleosome bound DNA). Number of ATAC-seq reads mapping to these windows from each condition was then quantified and used for the differential analysis (see section 2.12). This method allows increasing resolution of differential analysis of chromatin accessibility within the called peaks from both conditions, since it allows focusing on a part of a peak or a putative cis-regulatory module.

Through the use of this method 5761 loci within 37370 merged peaks from the 48hpf gut and whole embryo samples were found to be relatively more open in the gut tissue compared to the whole embryo (Figure 3.7). Conversely, this

method also identified 7204 loci that are relatively more closed, which could be useful for predicting CRMs that could drive expression of target genes in other domains, if the target gene of interest is expressed in other domains rather than the embryonic gut. *SpLox*, for instance, shows expression in the nerve cells of the sea urchin embryo (Cole & Arnone 2009), thus the less open regions of the ATAC-seq peaks in the gut samples could account for this ectodermal expression. Using the same method on the 66 hpf gut and whole embryo data sets, 2315 loci relatively more accessible in the gut were identified within 34380 merged ATAC-seq peaks from the two conditions, as well as 1541 loci were found to be less open in the gut sample (Figure 3.7). The lower number of differentially accessible regions was identified for 66 hpf due to replicate issues, which were already discussed, so it is likely that only the most contrasting loci were singled out between the two conditions.

High enrichment of motifs corresponding to the vertebrate homologs of transcription factors SpFoxA, SpBrn1/2/4, SpOtx and SpHnf1 was detected in these relatively more open gut ATAC-seq peak loci in both 48 hpf and 66 hpf datasets using de-novo motif predictions, performed by HOMER (Figure 3.7). All of these TFs are involved in endoderm specification leading to development of different parts of the gut (Smith et al. 2008; de-Leon & Davidson 2010; Yuh et al. 2005; Perillo et al. 2016). *SpFoxA* is expressed throughout the gut (Tu et al. 2006), SpBrn1/2/4 is expressed in the esophagus (Annunziata & Arnone 2014), while *SpHnf1* domain of expression is in the stomach (Perillo et al. 2016). *SpOtx* controls the development of the posterior gut at 48 hpf, expanding its expression to other digestive system parts by the pluteus stage, as well as in oral ectoderm

(Yuh et al. 2002). In addition, some transcription factor motifs associated with the coelomic pouches, such as SpSoxF (Luo & Su 2012) and SpFoxF (Tu et al. 2006), were also found enriched in this dataset (Figure 3.7). This is not surprising since the coelomic pouches are located at the tip of archenteron at the 48hpf gastula and near the foregut and midgut at the later stages of the sea urchin embryogenesis into the prism and pluteus stages (Luo & Su 2012). Found motif enrichments suggest that the differential analysis of the gut ATAC-seq datasets identified reasonable candidates for gut specific regions of the putative CRMs.

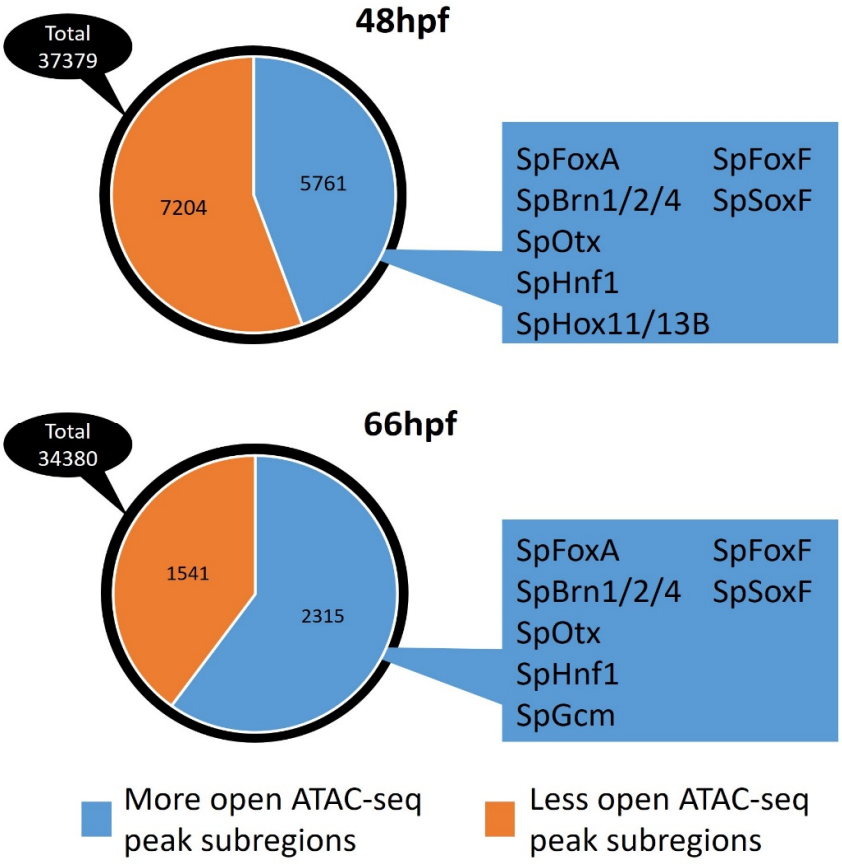


Figure 3.7 Gut enriched peaks. Pie charts showing the total number of gut peaks, proportions of more open and more closed regions in the differentially accessible peaks, as well as transcription factor motifs enriched in the more regions more open in the gut datasets.

3.2.6 ATAC-seq predicts putative CRMs around ParaHox genes Lox *and* Cdx

Obtaining the loci of open chromatin that could potentially be functional CRMs allows gaining insight into which genes these putative CRMs control. Using annotatePeaks.pl tool from HOMER (Heinz et al. 2010) (see section 2.13) nearest transcription start sites were identified for each open chromatin region. This allowed identifying 11 putative CRMs with the ParaHox genes as the potential targets in the sea urchin *S. purpuratus.* Of these, six putative CRMs were found likely to affect *SpLox,* while the remaining five affect *SpCdx.*

SpLox putative CRMs are distributed throughout the 30 kb gene model with one putative falling into the promoter/ transcription start site/ first exon region, and the rest falling within the only gene intron (Figure 3.8), notably, there are no peaks in the second (final) exon of the gene. In terms of genomic length, the majority of the putative CRMs are contributed by the ATAC-seq peaks at 48 hpf, especially from the gut sample. Of course, this reflects how the putative CRMs are defined, however, it also indicates that there are wider ATAC-seq peaks in the 48 hpf gut enriched dataset that are significant from the peak calling and replicate combination. This could be significant for gene regulation since *SpLox* expression starts at around 40 hpf and at 48 hpf it reaches its highest expression levels (Tu et al. 2014), so such a profile of open chromatin could account for the onset of regulation of the expression of this gene in the different regions of the sea urchin gastrula, since *SpLox* is expressed ectodermally as well as in the gut. Majority of the putative CRMs are open, at least in part, at most time points for which the

datasets were generated, except SpLoxCRM7 which is contributed solely by the whole embryo 48 hpf dataset, since in this dataset this peak is significant while in others it is not. Still, differential analysis of the putative CRM described in the previous section, shows that this locus is not differentially less open in the gut tissue, although in the gut tissue this peak is not significant from the peak calling and replicate combination. The putative CRM falling within the transcription start site (TSS), SpLoxCRM9, is open throughout the developmental stages for which the data was obtained. Differential analysis of this putative CRM shows that there is a locus that is less open in the gut enriched sample and also a region which is more open in the gut upstream of the TSS, along with another region in SpLoxCRM1-4 (Figure 3.8). This indicates the potential roles of these regions as regulatory elements in different sea urchin tissue types. This will be explored in the following chapters.

SpCdx putative CRMs are located in the upstream intergenic region, TSS and exons, including the last exon of the gene annotation, no intronic CRMs were identified for this gene (Figure 3.8). For this gene, the majority of genomic lengths of these CRMs are contributed by the 66hpf dataset, which, again, could indicate important regulations at the timepoints of highest expression (Arnone et al. 2006), although its expression starts earlier. Out of five putative CRM, only one contains a region which is more open in the gut tissue, while three different CRMs are actually less open in the gut samples, which is somewhat surprising, since this gene in the sea urchin embryo is exclusive to the gut. This could be due to repression of SpCdx through these CRMs in regions other than the gut (e.g. ectoderm) or related to the gut tissue collection method, as this is a posterior-

most gene, and some cells from the anus region of the gut tissue could have been lost during the mechanical separation of the tissue. In the latter case, only the most accessible genomic ATAC-seq loci would be detectable as differentially more open in the gut samples. The SpCdxCRM1 has a region that is comparatively more accessible in the gut compared to the whole embryo in both 48 hpf and 66 hpf datasets, suggesting that this peak is likely to drive expression of SpCdx in the gut. Accessibility of the chromatin around the genomic loci occupied by the ParaHox genes, therefore, gives an indication to what regions of the genome could act as CRMs and which of them could be gut specific.

Figure 3.8 *S. purpuratus* ParaHox putative CRMs near *SpLox*, *SpCdx*, *PlLox* and *PlCdx*.

Using the same method 16 putative CRMs were identified for the *P. lividus* ParaHox genes. Eight of these are identified around *PlLox* and another eight around *PlCdx*. The *PlLox* putative CRMs are located in the upstream region from the gene transcription start site, overlapping the TSS and in the introns, as well as one putative CRM in the region downstream of *PlLox*. Again, as in the case of *SpLox,* there are no open chromatin regions in the second (last) exon. The PlCdx putative CRMs are located upstream, overlapping the TSS, in the introns and in the third exon. It is worth noting that from the currently available annotation *PlCdx* overlaps another gene model: *PlAnL18*, which signifies that the last exon of *PlCdx* does not have putative CRMs, if this annotation is correct (Figure 3.9).

S. purpuratus and *P. lividus* show high degree of similarity in six CRMs in both species, three per each gut ParaHox gene (Figure 3.8). SpLoxCRM9 is similar to PlLoxCRM2 with over 84% sequence similarity, and SpCdxCRM4 and PlCdxCRM5 exhibit 82.5% identity. Similarity in these regions in the two closely related species is not surprising since these CRMs overlap the transcription start sites, which are likely to code for functional regions of the Lox protein. The exonic CRMs near *Cdx,* SpCdxCRM5 and PlCdxCRM8, are also conserved with 79.2% identity. The intronic putative CRMs from the two species are over 73% identical, which is lower than for the putative exonic CRMs, however considering that the intronic regions are not coding, such percentage of identity could indicate conserved functional cis-regulatory elements. This could also be true for the intergenic *Cdx* CRMs, which are over 89% similar between the two sea urchin species (Figure 3.8), however this shared CRM is broken in *P. lividus* since sequences from two PlCRMs (PlCdxCRM1 and PlCdxCRM2) align with the

SpCdxCRM1 (Figure 3.9). In *S. purpuratus* these alignments are adjacent while in *P. lividus* there is more DNA sequence between them. The fact that these sequences are accessible in the chromatin, despite breakage in *P. lividus*, points to its function as a CRM.

Figure 3.9 *P. lividus* ParaHox putative CRMs near *PlLox*, *PlCdx* and *SpCdx*

In the *Patiria miniata* species the ParaHox genes are located in a cluster (Annunziata et al. 2013), and associating ATAC-seq peaks with the most proximal gene allows prediction of *PmLox* and *PmCdx* putative CRMs (Figure 3.10). *PmLox* was found to be adjacent to 20 putative CRMs, most of which (17 ouf of 20) are in the likely intergenic region between *PmGsx* and *PmLox,* while the remaining three are in the first intron, the last exon and the region between *PmLox* and *PmCdx*. *PmCdx* was associated with eight putative CRMs by HOMER, four of which are between *PmLox* and *PmCdx,* one intronic CRM, one exonic and one overlapping the transcription start site. Noticeably, there are no ATAC-peaks overlapping the transcription start site of *PmLox.* However, this could be due to issues with gene annotation. In addition, the high number of peaks upstream of *PmLox* may not be representative of the real situation due to the assembly issues. *PmLox* and *PmGsx* should be around 30 kb apart from the BAC sequencing (Annunziata et al. 2013), while the current genome assembly implies that there is a gene free stretch of DNA of more than 40 kb from the beginning of *PmLox* gene model (also discussed in section 3.1) (Figure 3.10). This issue makes *PmLox* CRMs analysis difficult, emphasizing the need for a better genome assembly. Pairwise putative CRM sequence comparisons have identified only one region with a high similarity among the three echinoderm species (Table 3.4). Despite the short length of this region of only 54 bp (Table 3.4), its existence in the three species suggests its importance for regulation of *Cdx* genes, especially considering that this region is in non-coding regions upstream of the gene transcription start sites in all three species (Figure 3.10). In

the sea urchins it is in the intergenic CRMs (Figures 3.8 and 3.9), while in the sea star it is in the CRM overlapping the TSS (Figure 3.10).

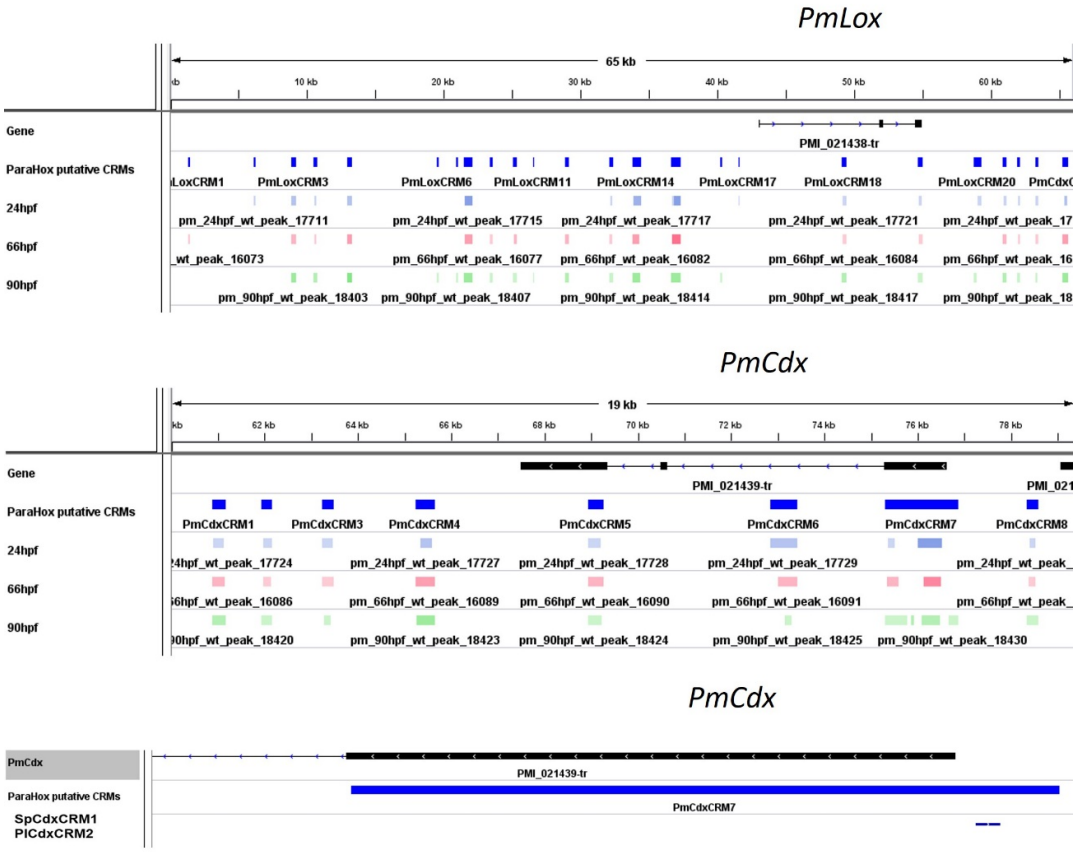


Figure 3.10 *P. miniata* ParaHox putative CRMs near *PmLox*, *PmCdx*, *PlLox*. Bottom panel shows SpCdxCRM1 and PlCdxCRM2 similar sequences near *PmCdx*

The *B. lanceolatum* ATAC-seq profile identified three peaks per each ParaHox gene (Figure 3.11). BlCdxCRMs are exonic with two of the three falling within the BlCdx gene body, while the third one, BlCdxCRM4, was identified inside the BlLox last exon. The proximity of these genes on the scaffold could lead to the misattribution of the putative CRMs if they are functional, and could lead to co-

regulation of different ParaHox genes by the same CRM, which could control collinearity. BlLoxCRMs are in the intron, overlapping the TSS and upstream of the gene model. BlGsxCRMs are also nearby due to tight gene clustering and are located upstream of *BlGsx*, at the TSS, and downstream of it (Figure 3.11).

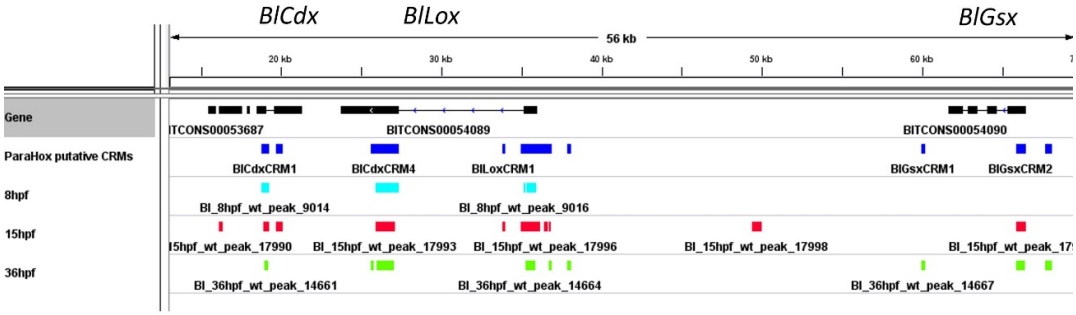


Figure 3.11 *B. lanceolatum* ParaHox putative CRMs near *BlCdx*, *BlLox* and *BlGsx*.

Pairwise sequence comparisons did not find large conserved regions between the *B. lanceolatum* putative ParaHox CRM regions and the ParaHox CRM regions from other species, aligning with multiple regions in the putative ParaHox CRMs belonging to non-corresponding genes (BlGsxCRM2 and SpLoxCRM1-4, for instance) and even with regions from the same species but from different putative CRMs (Table 3.4). This could suggest conservation between the sequences in regulatory regions between species that have an intact cluster and those that don't, but due to their small size (less than 30 bp) and sequence repetitiveness these hits are not informative as to sequence conservation between the amphioxus and echinoderm putative ParaHox CRMs.

Table 3.4 Examples of putative CRM sequence similarities obtained via pair-wise BLAST searches with low scores

Query CRM ID	Subject CRM ID	Percentage identity	Score	Aligned sequence
PmLoxCRM3	SpLoxCRM1-4	100	22	TGATATGGCTC
BlGsxCRM2	PlCdxCRM5	93.333	24	GGCGAAGGTAAACAA
BlGsxCRM2	BlLoxCRM2	84.211	22	CTTTTCTTTCGCTTTATCT
PmLoxCRM4	PmLoxCRM18	100	26	GCTTCCCGTCCGC
PmCdxCRM7	SpCdxCRM1	74.074	34	GAATTTTATGACTCTTTCTAAACAATTTGCAGCAACGGCGTATTGTTTGCGTTG
PmCdxCRM7	PlCdxCRM2	74.074	34	GAATTTTATGACTCTTTCTAAACAATTTGCAGCAACGGCGTACTGTTTGCGTTG
PlCdxCRM2	PmCdxCRM7	74.074	34	GAATTTTACGACT-TGTCAAGATGAGCTTGAGCAACGGGCTAGTGTTTGCGTTG
SpCdxCRM1	PmCdxCRM7	74.074	34	GAATTTTACGACT-TGTCAAGATGAGCTTGAGCAACGGGCTAGTGTTTGCGTTG

There are peaks adjacent to the putative CRMs described, however they are closer to the transcription start sites of other genes, flanking the ParaHox loci. Indeed, these open regions may also play a role in the control of the ParaHox genes. Yet these were excluded from the current analysis since, to author's knowledge, there is no intrinsic information in the data available that suggests a non-biased way of predicting which of these ATAC-seq peaks, most proximal to other genes, could control the ParaHox genes. HiC data, currently, does not have enough resolution to predict elements that are very close to each other in cis (less

than 20kb apart), since, in such cases, the high read count identifying three-dimensional interactions between the loci could be due to the linear proximity. Therefore, the work performed during the thesis focuses solely on peaks whose most proximal genes are ParaHox genes, however, in the future more distal putative regulatory elements of these genes will be also be addressed.

3.3 Discussion

Assessment of the ParaHox gene organization within the genome and identification of open chromatin regions allows to predict the cis level of regulation of expression of these genes. In the sea urchin species, the genes are far apart from each other on the single chromosome and have no significant three-dimensional interactions with each other. Such remoteness suggests that, in case of the sea urchins, the ParaHox genes are unlikely to be co-regulated in cis. The putative CRMs identified through bioinformatic analysis of open chromatin data obtained via ATAC-seq experiments are reasonable candidates, since the same approach used to predict ParaHox CRMs, is capable of identifying known CRMs when used on the published cis-regulatory elements. Differential analysis of open chromatin also gives an indication to which regions of the putative CRMs contribute to gene regulation in the specific tissues, in this particular case, the embryonic gut. Conservation between the putative CRMs of the two sea urchin species, compared to other taxa examined (sea star and amphioxus), suggests that the regulation of the ParaHox genes in them could happen through comparable genomic regions and, potentially, through the same transcription factors, since the cis-regulatory elements are functional through interaction with

these trans-acting agents. Therefore, it is essential to identify the transcription factors that are capable of binding the identified CRMs in the cell or tissue type of interest and at a specific time point, in order to shed light on the regulation of the ParaHox genes as well as the evolution of this regulation.

CHAPTER 4

IN SILICO GRN DRAFTING THROUGH COMBINING -OMICS DATA

This chapter describes an *in silico* approach to uncovering GRN topologies, using the different transcriptomic data sets (differential RNA-seq and scRNA-seq) along with the open chromatin ATAC-seq datasets, which were used to identify the putative CRMs. Results pertaining to the scRNA-seq and the differential RNA-seq data discussed in this chapter can be found in the Non-book component files on the USB drive.

4.1 Introduction

Gene expression control in various tissues and cell types at different time points can be visualized through the use of gene regulatory networks. Two of the main components of a GRN are nodes and their interactions. Nodes are genes, while their effect on each other via transcribed RNA and translated proteins constitutes interactions between the nodes. The transcription factors physically bind specific DNA sequences to facilitate control of the target gene expression by bringing the transcription machinery to the promoter of a target gene (Ong & Corces 2011). If these regions are located on the same chromosome as the target genes they are called cis-regulatory regions or modules (CRMs). The approach described in the previous chapters allowed prediction of a number of putative proximal CRMs for the two ParaHox genes involved in the gut development in the four deuterostome species. In order to draft a GRN, the transcription factors that bind these CRMs need to be identified. Components and wiring of the GRNs from different species can be used to gain insight into the evolution of gene control, cell type differentiation and embryogenesis in general.

It is possible to predict these transcription factors bioinformatically by searching for transcription factor binding sites within the genomic regions of interest. Software tools like HOMER use motif position weight matrices in order to identify sequences where a transcription factor is able to bind within a given set of genomic loci, such as putative CRMs predicted by the ATAC-seq data analysis, described in the previous chapter, as well as perform *de novo* motif prediction and provide motif enrichment information for these loci (Heinz et al. 2010). If the

putative CRMs chosen affect genes that code for transcription factors, then position weight matrices for these transcription factors could then be used to predict where these TFs can bind within given genomic loci (such as putative CRMs). Therefore, data and analysis allows to predict the nodes of the GRNs and their interactions (Lowe et al. 2019).

However, transcription factors as well as CRMs can be tissue type and time-specific (Ong & Corces 2011). Therefore, expression of transcription factors that can bind the CRMs of interest at the given time point and tissue or cell type needs to be assessed. To this end, transcriptomic data needs to be obtained, in particular, the gut enriched RNA-seq to assess expression of transcription factors in the gut tissue or cell specific transcriptomic methods such as single-cell RNA sequencing (scRNA-seq). These methods allow identification of differential transcription factor expression in the time and space specific manner. As the name implies, scRNA-seq allows gene expression assessment at a single cell resolution by using cell specific barcodes to assign cell identities in post-sequencing analysis, allowing identification of cell clusters, cells of which show similar transcriptomic profiles that could be cell or tissue types (Hwang et al. 2018). Combining these transcriptomic datasets with CRM predictions from the ATAC-seq allows GRN drafting (Lowe et al. 2019; Lowe et al. 2017). In this chapter, a draft of the GRN upstream of *Lox* and *Cdx* genes in *S. purpuratus* at 72hpf, built through this combinatorial approach, is presented.

4.2 Results

4.2.1 Predicted binding of transcription factors within putative CRMs

The loci that were identified to be putative CRMs should contain transcription factor binding sites within them. These binding sites were predicted through the use of motif matching tools and the position weight matrices (PWMs) for these motifs from JASPAR2018 database (Heinz et al. 2010; Khan et al. 2018). JASPAR2018 contains 579 PWMs for vertebrate transcription factors. The putative CRMs for *S. purpuratus* were found to contain motifs for 351 vertebrate transcription factors, with 181 of these motifs found in the *S. purpuratus* CRMs more than once. In the *P. lividus* CRMs 396 vertebrate transcription factor binding sites were detected, again, with as much as 239 of them appearing more than once. In the *P. miniata* putative CRMs even more transcription factors could possibly bind since 501 PWMs of 579 were found in these loci, with 360 more than once. The amphioxus CRMs motif matching analysis also resulted in the detection of a high number of possible TF binding sequences as 465 TFs were found to potentially recognize sequences within these CRMs. Of these, 316 were non unique in the CRMs.

Comparing how many motifs could be shared between the different species 225 were found to be shared within all ParaHox putative CRMs among the four species. *P. miniata* and *B. lanceolatum* share the most transcription factor binding sites, which could be related to the fact that the ParaHox genes are in a cluster, but it is most likely to be due to the sheer numbers of the putative TF binding sites detected, since in these two species more of the vertebrate PWMs were found

(Figure 4.1 A2). This is followed by the sea urchin species sharing 303 putative TF binding sites. The echinoderms share 265 putative transcription factor binding sites (Figure 4.1 A2). Here it is worth mentioning that only 3 TF motifs from the 579 vertebrate PWMs were not found in any of the parahox CRMs in any of the species examined.

Lox putative CRMs showed 66 motifs shared among all species, with most shared transcription factor binding sites between the echinoderm species. *S. purpuratus* and *P. lividus* share 197 different TFs, while all three echinoderms share 144 TFs. 48 were found not be in any of the putative *Lox* CRMs in any of the speceis (Figure 4.1 B2).

Simiar situation is visible for the *Cdx* putative CRMs, however the number of the TFs that could bind withing these CRMs is lower than for *Lox* genes. Only 33 TF motifs are shared among all four species (Figure 4.1 C2), with *P. miniata* and *B. lanceolatum* sharing most TF motifs. In case of *Cdx,* there are more non-shared transcription factor binding sites with 16 being uniquely found in *S. purpuratus,* 20 in *P. lividus,* 96 in *P. miniata* and 65 in *B. lanceolatum*. 105 TFs from 579 were found unable to bind any of the *Cdx* putative CRMs since their motifs were not found (Figure 4.1 C3). Consistently *S. purpuratus* and *P. lividus* share high number of TFs that can bind within their CRMs based on binding site predictions from genomic sequences and the position weight matrices, this is possibly due to the CRM sequence conservation between the species (Figure 3.8).

High numbers of motifs were found within the putative CRMs more than once. The motif occurrence counts were used for principal component analysis (PCA)

(section 2.14). The PCA graphs for the motif occurrences within all the ParaHox gene CRMs found that the two sea urchin species cluster together, while the sea star and amphioxus are further away (Figure 4.1 A1). The same tendency is noticeable in the *Lox* putative CRMs and the *Cdx* putative CRMs: the sea urchin motif occurrences make them cluster together, while *P. miniata* and B. *lanceolatum* are separate (Figure 4.1 B1 and C1). Some of the variation between the PmLoxCRMs and other echinoderm *Lox* CRMs could be due to the assembly issues discussed previously. Nonetheless, the principal component 1 (PC1) describes the most variation in the data compared to the principal component 2 (PC2) (Love et al. 2014). This indicates, that *P. miniata*, although far from the sea urchins on the PCA plot, has more similarity to them in terms of motif occurrence than *B. lanceolatum* has*,* since it is closer to them on the PC1 axis (Figures 4.1 A1, B1 and C1). The motif occurrences are also likely to depend on the overall length of the CRM regions, however, unlike simple presence or absence of a motif, it is more sequence conservation dependent. These results indicate that the motif occurrence is dependent on the evolutionary closeness of the species, since the sea urchins are the most closely related, with sea star being their next relative and the amphioxus being the most distant taxon in this case (Blair & Hedges 2005), and the PCA plots show the same pattern (Figures 4.1 A1, B1 and C1). Of course, to confirm this suggestion, motif occurrences within ParaHox genes of more taxa is required. The clustering of the species in terms of motif occurrence could be explained by different wiring of the GRNs upstream of the ParaHox genes. In order to test this notion, the GRNs need to be drafted, validated and compared: following sections and chapters describe the work done

for the GRN elucidation and evolutionary comparison so far, during the course of this thesis work.

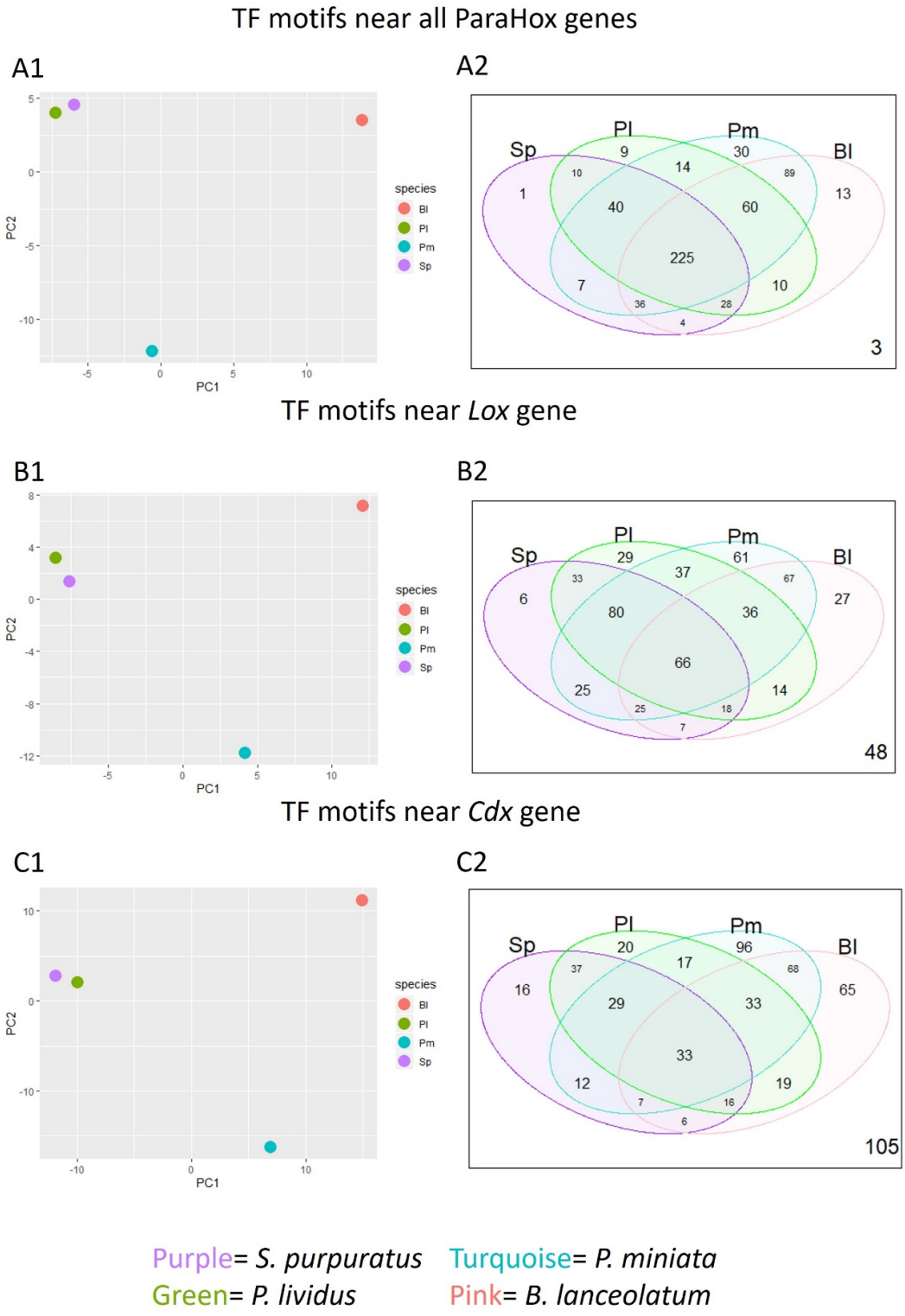


Figure 4.1 Predicted TF motif bound within ParaHox CRMs analysis. Panel A shows PCR plots for motif occurrence assessment, Panel B shows Venn diagrams indicating shared and unique TF motifs between the species of interest.

This approach for finding putative transcription factors is indeed applicable as shown by the analysis of the known CRMs.The authors of the corresponding papers, discussed in section 3.2.4, also showed direct effects of some of the transcription factors on the target genes by TF binding site predictions and mutagenesis. From the literature, it was possible to mark that *SpHox11*/13B was shown to be affected by itself and SpTcf (Cui et al. 2017a), *SpTbrain* was shown to be regulated by SpEts1/2, SpElk and SpEts (Wahl et al. 2009), *SpOtx* is directly activated by itself, SpGatae and SpBlimp1 (Yuh et al. 2004), *SpFoxA* is known to be regulated by SpBra (de-Leon & Davidson 2010), while SpBlimp1 regulated SpWnt8 expression (Minokawa et al. 2005). Using the HOMER software suite, every mentioned TF motif in the CRMs for their respective target genes could be confirmed, and more putative transcription factors that could bind these known CRM could be identified (see Non-book component).

This illustrates the predictive power of the approach based on the combination of various omics datasets and bioinformatical tools such as HOMER. Nonetheless, high numbers of predicted motifs both in the case of known CRMs and the putative ParaHox CRMs indicates that such motif matching tools give an overestimation of the transcription factors controlling a given CRM, leading to a need in filtering the predicted motifs. For instance, to filter for a given tissue and at a given time point transcriptomic data can be employed (Lowe et al. 2019). Such data transcriptomic data was obtained for *S. purpuratus*.

4.2.2 Gut enriched RNA-seq data reveals differentially expressed gut tissue genes

Gut tissue RNA libraries were obtained for the *S. purpuratus* 48 hpf gastrula and the 66 hpf prism the same way as for the ATAC-seq gut libraries described previously (section 2.2 and 3.2.5), except three biological replicates, instead of two, were made for the RNA-seq (Table 4.1). After sequencing, the resulting reads were pseudo-aligned onto the *S. purpuratus* transcriptome to obtain counts per each transcript and perform differential analysis (section 2.16).

Table 4.1 RNA-seq datasets information used to identify gut specific genes *S. purpuratus*.

S. purpuratus **RNA-seq**			
Sample	Number of reads	Number of reads after trimming	Alignment rate
48hpf whole embryo 1	16575251	15328866	71.96%
48hpf whole embryo 2	18035777	16304323	71.37%
48hpf whole embryo 3	18711339	17477008	70.54%
48hpf gut tissue 1	13028246	11153272	63.54%
48hpf gut tissue 2	18371321	10723587	57.68%
48hpf gut tissue 3	17293473	15453098	59.62%
66hpf whole embryo 1	30264023	28100526	68.99%
66hpf whole embryo 2	30306638	28225474	68.38%
66hpf whole embryo 3	28608881	26487858	67.80%
66hpf gut tissue 1	16581104	14458830	63.56%
66hpf gut tissue 2	33630233	29345187	68.32%
66hpf gut tissue 3	12343381	10841418	64.00%

A set of 3929 transcripts was identified as differentially expressed in the gut samples compared to the whole embryo samples at 48 hours post fertilization using the p-adjusted value of 0.05 as the cut off. Upregulated transcripts represented 38.7% of the total differentially expressed transcripts, as 1524 of

them were identified (Figure 4.2). At the 66 hpf prism, 1489 transcripts were identified to be expressed in the gut differently than in the whole embryo, with 400 of them being upregulated (Figure 4.2). Upregulated transcripts are the most interesting in this case, since they are the ones that are present in the gut in higher amounts compared to the rest of the embryo and, therefore, are more likely to contribute to the gut GRNs. The identified transcripts were then combined with the ATAC-seq data, to test the quality of the resulting RNA-seq data, by identifying which of the upregulated genes are adjacent to the differentially more accessible genomic loci identified via the ATAC-seq peaks from the corresponding time points. 487 of the 1524 genes upregulated in the gut at 48 hpf also have ATAC-seq peaks nearby that are more open in the gut. (Figure 4.2). At the 66 hpf 193 of 400 are upregulated and have more open ATAC-seq peaks near them in the gut tissue samples (Figure 4.2). The number of genes with higher expression in the gut datasets that have differentially more open chromatin in the gut datasets is likely to have a biological significance since the random 150 basepairs peak sets were proximal to significantly lower number of upregulated gut genes (48 hpf p-value= 1.3801248414031704e-30, 66 hpf p-value= 1.3916508986397104e-201). 84 of these genes are more expressed in the gut samples and have putative CRMs with the regions that are more open in the gut at both time points. 225 are only found at 48 hpf, while 56 are specific to 66 hpf (Figure 4.2). Shared genes include *SpBlimp1*, *SpFoxA* and *SpTgif*, all of which are known to be involved in the gut GRN, with SpBlimp1, which as known activator of *SpLox*, (Livi & Davidson 2006; Smith et al. 2008) and *SpFoxA* (Tu et al. 2006; de-Leon & Davidson 2010) specifying the endoderm from the early

developmental stages and *SpTgif* being expressed around the blastopore and in the coelomic pouches (Howard-Ashby et al. 2006). Notable differentially more expressed genes in the gut at 48 hours include *SpLox, SpEndo16, SpNkx6.1, SpGatae, SpHnf4, SpNos1* and *SpPla2* (Figure 4.2). All these genes have also been shown to be expressed in the different regions of the gut: *SpLox*, which is one of the genes of interest for this thesis, is expressed in the hindgut, in particular after 66hpf in the pyloric sphincter and the intestine, *SpEndo16* is a gut terminal differentiation gene, while *SpGatae* regulates the endoderm specification and the gut development as mentioned previously. *SpNkx6.1* was identified as a pancreatic gene (Annunziata et al. 2014), pancreatic type cells are located in the stomach of the sea urchin embryo (Perillo et al. 2018). *SpHnf4* is a likely homolog of the vertebrate HNF4A gene, which specifies the endoderm and is related to the digestive system diseases such as diabetes (Duncan et al. 1994). *SpNos1* has been suggested to control sphincter contraction between gut compartments of the sea urchin embryo, in addition to a neuronal role (Yaguchi & Yaguchi 2019). *SpPla2* is also likely to be an important gene for the gut development in the sea urchins since its likely homolog was shown to pattern adult starfish pyloric cecae (Kishimura & Hayashi 2005). This shows that, indeed, the gut tissue differential RNA-seq identifies some genes involved in the embryonic gut patterning or function. However, neither *SpCdx* nor *SpHox11/13b* were found differentially expressed in these gut samples compared to the wild-type embryo, despite them being mostly gut specific. In addition, these genes were also absent from the temporal differential gut RNA-seq analysis (see Non-book component). These genes are posterior-most genes expressed in the

blastopore/anal region of the gut, and therefore it is possible that some of the anal cells from the gut tissue were lost. Presence of *SpTgif* in the differential list, despite it also being expressed in the anus, is possibly due to its expression in the coelomic pouches (Howard-Ashby et al. 2006). It is also likely that the gut tissue collected had coelomic pouches due to their adjacency on the embryo and due to the fact that coelomic pouch genes were also found to be differentially more expressed in the gut samples. At 48 hpf these genes include *SpFoxC*, *SpFoxY* and *SpScratchX*, while at 66 hpf genes like *SpScl* were detected (Figure 4.2), all of which have expression in the coelomic pouches (Materna et al. 2013; Tu et al. 2006; Solek et al. 2013).

The RNA-seq results indicate that the gut samples seem to have coelomic pouch tissues within them as suggested by the differential ATAC-seq data as well. Again, due to the way such data is obtained the gut datasets are simply gut cell enriched compared to the whole embryo. In addition, while it is possible to identify the differentially expressed genes, these genes are not the only ones that could have important roles in the hindgut GRNs. This highlights the need for higher resolution transcriptomic data.

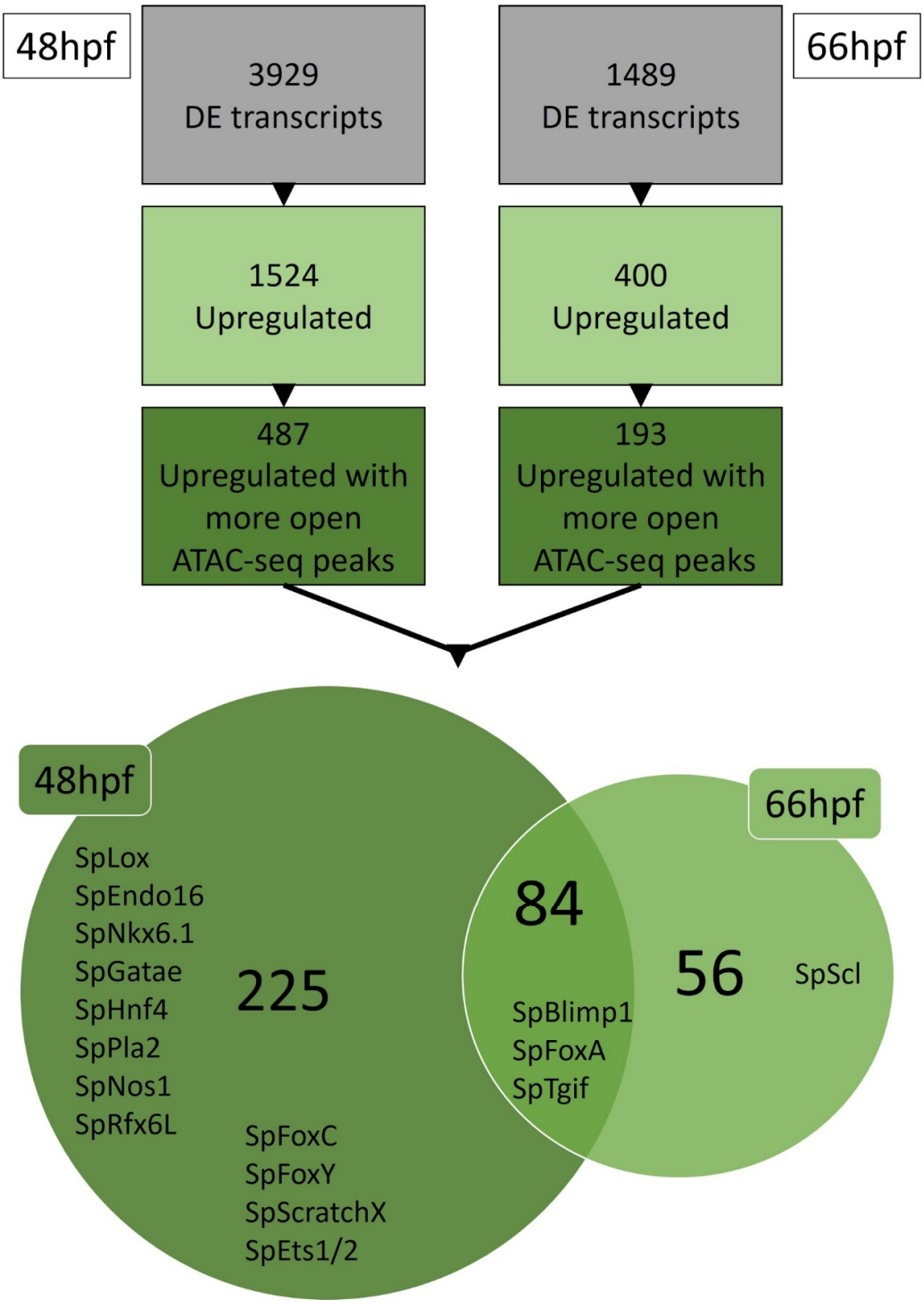


Figure 4.2 Transcripts differentially expressed in the gut samples compared to whole embryo at 48 and 66 hpf. Total differentially expressed transcript number is in grey, light green shows how many of total are upregulated in the gut, dark green shows how many of upregulated are close to

an ATAC-peak regions that are more open in the gut. Venn diagram shows the differences and similarities in the make-up of the upregulated transcripts with more open peaks at 48 and 66 hpf.

4.2.3 Single cell RNA-seq identifies cell populations, belonging to hindgut regions, and their transcriptomic profiles

Single cell RNA-seq data allows cell resolution identification of transcripts, which allows grouping of cells into clusters based on the similarities of their transcriptional profiles. These clusters correspond to cell populations, cell types or even tissues. The single cell approach was used to identify clusters which have the ParaHox genes in their transcriptional profiles in collaboration with Periklis Paganos, a PhD student at the lab of Dr Arnone at the time of collaboration, and Jacob Musser from Dr Detlev Arendt's lab. The approach allowed to obtain transcriptomes of 30000 cells from 6 datasets with all samples having median of unique molecular identifiers (UMI), representing individual mRNA, of more than 200 (Table 4.2).

Table 4.2 scRNA-seq datasets information used to identify cell clusters in *S. purpuratus* at 72hpf. E indicates extra-sequencing of the same sample.

S. purpuratus **scRNA-seq**					
Sample	Number of cells	Number of reads	Alignment rate	Number of reads per cell	Median UMI Counts per Cell
72 hpf 1	4000	26473738	61.10%	6618	243
72 hpf 1 E	4000	88568769	76.10%	22142	645
72 hpf 2	7000	51602660	69.30%	7371	409
72 hpf 2 E	5000	160618439	75.50%	32123	1329
72 hpf 3	5000	107226457	78.80%	21445	856
72 hpf 4	5000	160618439	75.50%	32123	1329

The single cell data of *S. purpuratus* 72 hpf plutei allowed identification of 21 clusters from 19699 cells, since 10301 cells were discarded due to low UMI counts during analysis. Mesoderm derivative cells including immune system cells, skeletal cells (Figure 4.3 Clusters 1 and 4), muscles (Figure 4.3 Cluster 19) and coelomic pouches (Figure 4.3 Clusters 18 and 20). Ectoderm clusters were also identified which contain neurons (Figure 4.3 Cluster 17) and other ectoderm derived cells (Figure 4.3 Clusters 2, 5-7, 14). Endodermal germ layer derivatives are represented by esophagus (Figure 4.3 Cluster 10), stomach cell clusters (Figure 4.3 Clusters 3, 7, 12 and 15), constituting different cell populations of the sea urchin larval stomach, pyloric sphincter cells (Figure 4.3 Cluster 11), intestine cells (Figure 4.3 Cluster 8) and anus (Figure 4.3 Cluster 9). Identity of the clusters was recognized by identifying the cluster marker genes and recognizing known cell and tissue type markers from this list, or by fluorescent RNA *in situ* hybridization experiments (FISH) performed by Periklis Paganos. Full descriptions of every cell cluster, their markers, as well as regulatory states will be published. Here, the majority of clusters are not described in detail since they are less relevant, in the context of this thesis, for the GRN upstream of the ParaHox genes in the gut.

The gene expression information was obtained for all genes expressed in the cells of the 21 clusters (see section 2.18), allowing to extract transcription factor lists for each hindgut cluster via use of PFAM (Finn et al. 2014), BLAST searches (Altschul et al. 1990; Apweiler et al. 2004) and functional annotation available at Echinobase.org (Kudtarkar & Cameron 2017).

The ParaHox genes that control gut patterning, *SpLox* and *SpCdx,* are found in three clusters of the hindgut of the sea urchin embryo. *SpLox* is expressed in the pyloric sphincter and, in a slightly lower relative percentage of cells, in the intestine. *SpCdx* is also expressed in the intestine, however a higher relative percentage of cells express it also in the anus (Figure 4.3). This scRNA-seq information is also supported by the visualization of the expression of these two genes in the same hindgut regions by fluorescent *in situ* hybridizations (Figure 4.3), as well as by published works (Arnone et al. 2006; Cole et al. 2009; Annunziata & Arnone 2014; Annunziata et al. 2014).

Consequently, the transcription factors from these hindgut clusters are of most interest for the GRNs controlling *SpLox* and *SpCdx*. Therefore, all the transcription factors that show average expression of over 0.5 were identified. The pyloric sphincter has 88 transcription factors above this cut off, within the intestine cluster 99 transcription factors could be identified, while the anus, the

most posterior of the three regions, contains 89 transcription factors. Around 30% of TFs are shared pairwise (Figure 4.4).

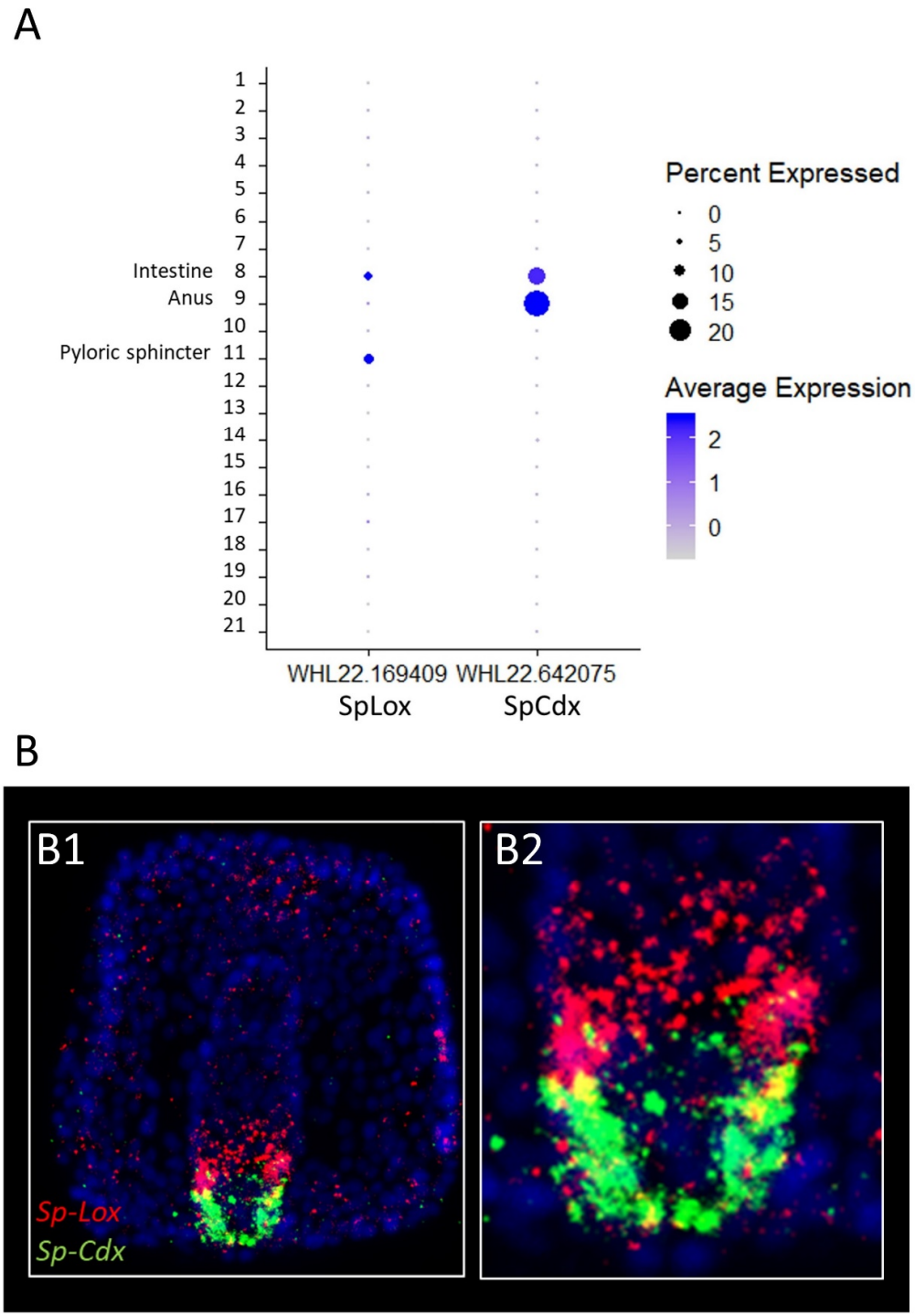


Figure 4.3 *SpLox* and *SpCdx* in the 72 hpf pluteus A. DotPlot of *SpLox* and *SpCdx* showing their expression in the hindgut clusters at 72 hpf. B. *In situ* images of *SpLox* and *SpCdx* expression in the 72 hpf pluteus. Courtesy of Periklis Paganos. B1. Full embryo. B2. Zoom-in on the hindgut.

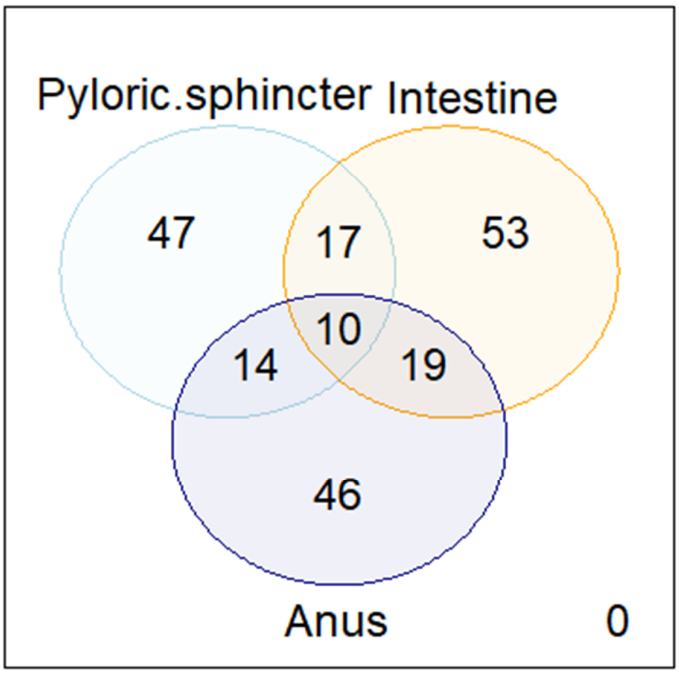


Figure 4.4 Venn diagram showing shared and unique transcription factors expressed in the three clusters of the hindgut: Pyloric sphincter, Intestine and Anus.

There are 10 transcription factors shared between every cell cluster (Figure 4.4). They *are SpEea1, SpNkrfL, SpFoxI, SpHox6, SpSmyd3_2, SpRfxc1l, SpRfx6L, SpMgam_3, SPU_025283/SPU_025284* and *SpE2f4*. Spatial expression patterns for most of these are unknown, however *SpFoxI* was shown to be expressed in the archenteron at 48 hpf and then in the hindgut regions at 72 hpf (Tu et al. 2006). Sea cucumber *SpHox6* homolog is known to function during gut regeneration. Expression of *SpRfx6L* was tested as part of cluster marker characterization via FISH and it was found to be expressed throughout the posterior endoderm regions, especially in the anus (Figure 4.6 E).

Top thirty most expressed transcription factors from the three embryonic hindgut regions show a number of known genes that pattern the gut and regulate its formation (Figure 4.5). These include *SpGatae* (Yuh et al. 2004), *SpHnf1aL* (same as *SpHnf1*) (Howard-Ashby et al. 2006), *SpNkx6.1* (Figure 4.6 A) (Annunziata & Arnone 2014), *SpFoxP* (Tu et al. 2006) and *SpLox* (Arnone et al. 2006) from the pyloric sphincter (Figure 4.5 A). The intestinal cells express *SpBlimp1* (Smith et al. 2008), *SpCdx*, *SpLox* (Arnone et al. 2006) and *SpFoxD* (Tu et al. 2006), which have been shown to have their spatial domains of expression in the posterior gut regions (Figure 4.5 B). Anal cells share some highly expressed TFs with the other two regions such as *SpHnf1* (Howard-Ashby et al. 2006), *SpBlimp1* (Smith et al. 2008), *SpCdx* (Arnone et al. 2006), as well as express *SpFoxA*, *SpFoxI* (Tu et al. 2006), *SpHox11/13b* (Howard-Ashby et al. 2006) and *SpBra* (Figure 4.5 C). Known expression patterns of these genes and visualization of expression of genes like *SpCdx* (Figure 4.6B), *SpHb9* (Figure 4.6 C) and *SpRfx6L* (Figure 4.6 E) via FISH support identification of the cluster expressing them as anal cells. Of course multiple of these genes such as *SpLox*, *SpCdx* (Figure 4.3 B), *SpFoxA* (Figure 4.6B) are shared between the three clusters even though they are not always in the thirty most expressed genes in each cluster.

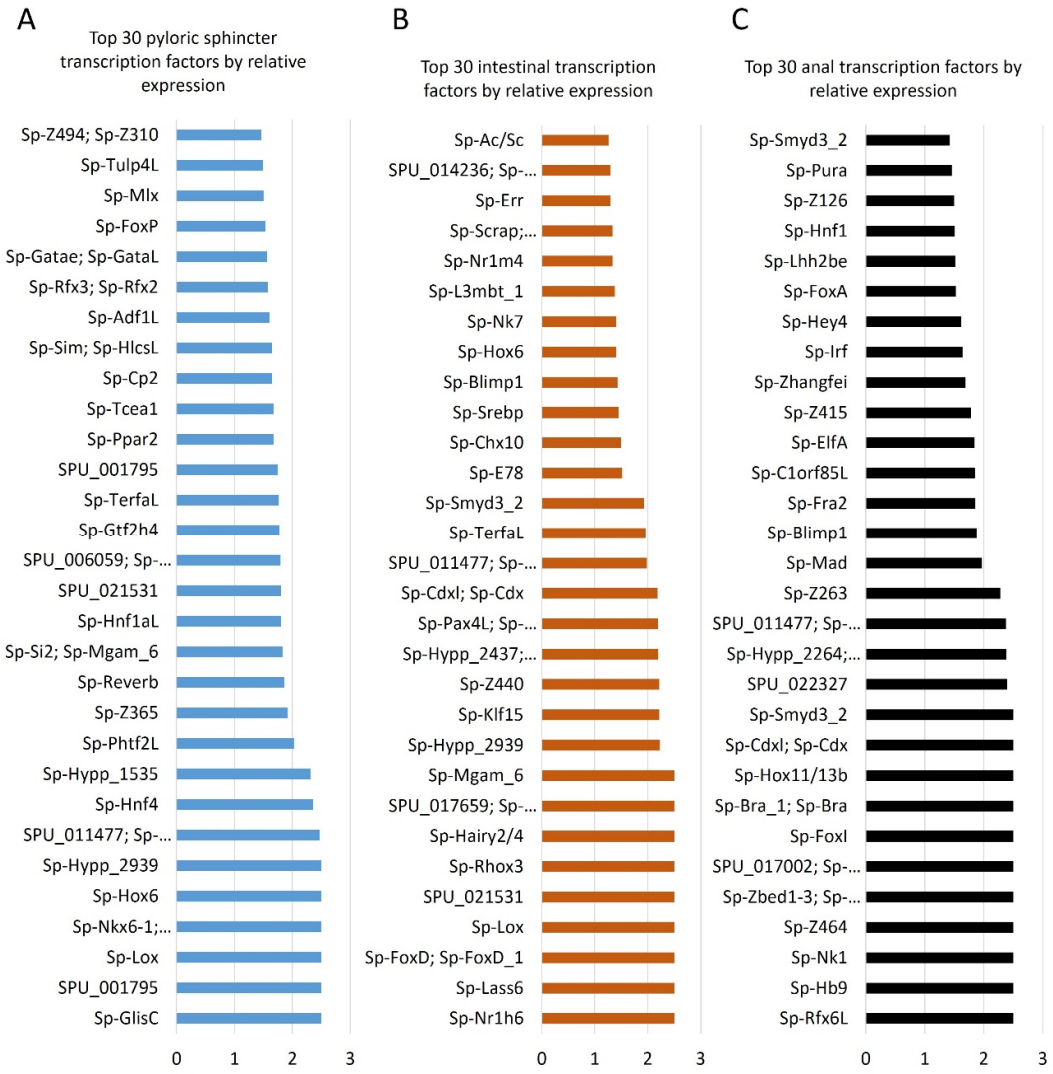


Figure 4.5 Most expressed transcription factors. A. Top 30 most expressed transcription factors in pyloric sphincter. B. Top 30 most expressed transcription factors in intestine. C. Top 30 most expressed transcription factors in anus.

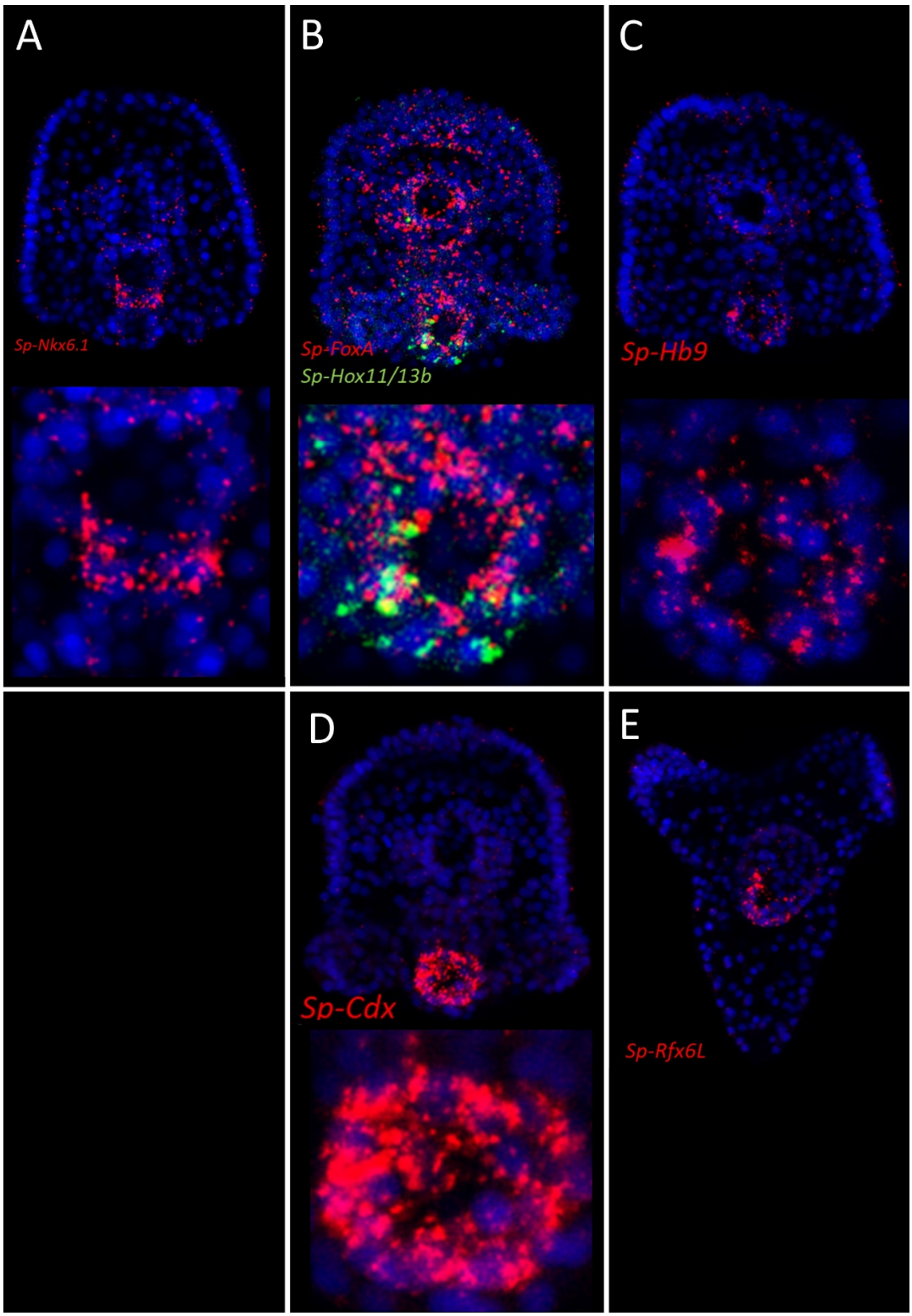


Figure 4.6 FISH images of markers of hindgut clusters at 72 hpf. Courtesy of Periklis Paganos. A. Expression of *SpNkx6.1* in the pyloric sphincter. B. Expression of *SpFoxA* (red) and

SpHox11/13b (green) in the anus. C. Expression of SpHb9 in the intestine. D. Expression of *SpCdx* in the anus and the intestine. E. Expression of *SpRfx6L* in the intestine.

This suggests that the cluster identities were recognised correctly, with scRNA-seq giving cell and tissue level resolution of expression, and that transcription factors from them can be used for drafting the GRNs controlling *SpLox* and *SpCdx*.

4.2.4 Draft GRNs controlling SpLox and SpCdx expression in the sea urchin embryonic gut

Identifying the cell clusters corresponding to the different regions of the hindgut allowed to pinpoint the transcription factors expressed in those regions. This information can be used to filter out what transcription factors, out of the ones that were predicted though the use of the motif matching software, are actually present in the cells that express *SpLox* and *SpCdx* at 72 hpf.

At 72 hpf, 32 unique transcription factors are co-expressed in those same regions as the ParaHox genes and can potentially bind the *SpLox* putative CRMs, nine of these have more than one potential binding site. SpLoxCRM1-4, SpLoxCRM5 and SpLoxCRM9 have the most transcription factor motifs identified. Clustering of TF binding sites is important for regulatory activity of a CRM (Peter & Davidson 2015). In the case of *SpLox,* SpLoxCRM5 and SpLoxCRM9 show regions where transcription factor binding sites are tightly clustered (Figure 4.7). Within SpLoxCRM5 this clustering could account for the *SpLox* self-regulatory loop, since this CRM contains the binding sites for PDX1, the likely homolog of *SpLox.* Another tight clustering of the transcription factor sites is adjacent to or within the

SpLoxCRM9 region, which is upregulated in the gut. This region contains a motifs for SpHox11/13b, SpCdx overlapping this motif, for SpHb9 and SpLmx1 (Figure 4.7). SpCdx represses expression of *SpLox* in the posterior-most section of the intestine (Annunziata et al. 2014; Annunziata & Arnone 2014) and the possibility of direct binding of SpCdx protein is evidenced by the data obtained during this thesis (Figure 4.7). Considering the known actors of the gut GRN can potentially bind within the region more accessible in the gut, SpLoxCRM9 is a promising candidate for a *SpLox* CRM. Transcription factors bound in the less accessible regions DNA regions could be activators of the CRMs that drive SpLox expression in the ectoderm or, in addition to that, repress SpLox in the gut. SpBra motifs found in the CRMs near *SpLox* (Figure 4.7) could also play a function in repressing the ParaHox gene since *SpBra* and *SpLox* do not co-express at 72 hpf. It is also worth mentioning that, in some cases, identification of the transcription factor targeting the CRM is impeded by similarity of JASPAR motif PWMs between the transcription factors. This is supported by motifs for SpLimc1, SpLmx1 and SpRox3 overlapping (Figure 4.7), due to them all being homeobox factors with similar DNA sequence recognition (Khan et al. 2018; Howard-Ashby et al. 2006). Existence of SpBlimp1 binding site suggests a direct interaction between the known gut activator SpBlimp1 and *SpLox* (Annunziata & Arnone 2014).

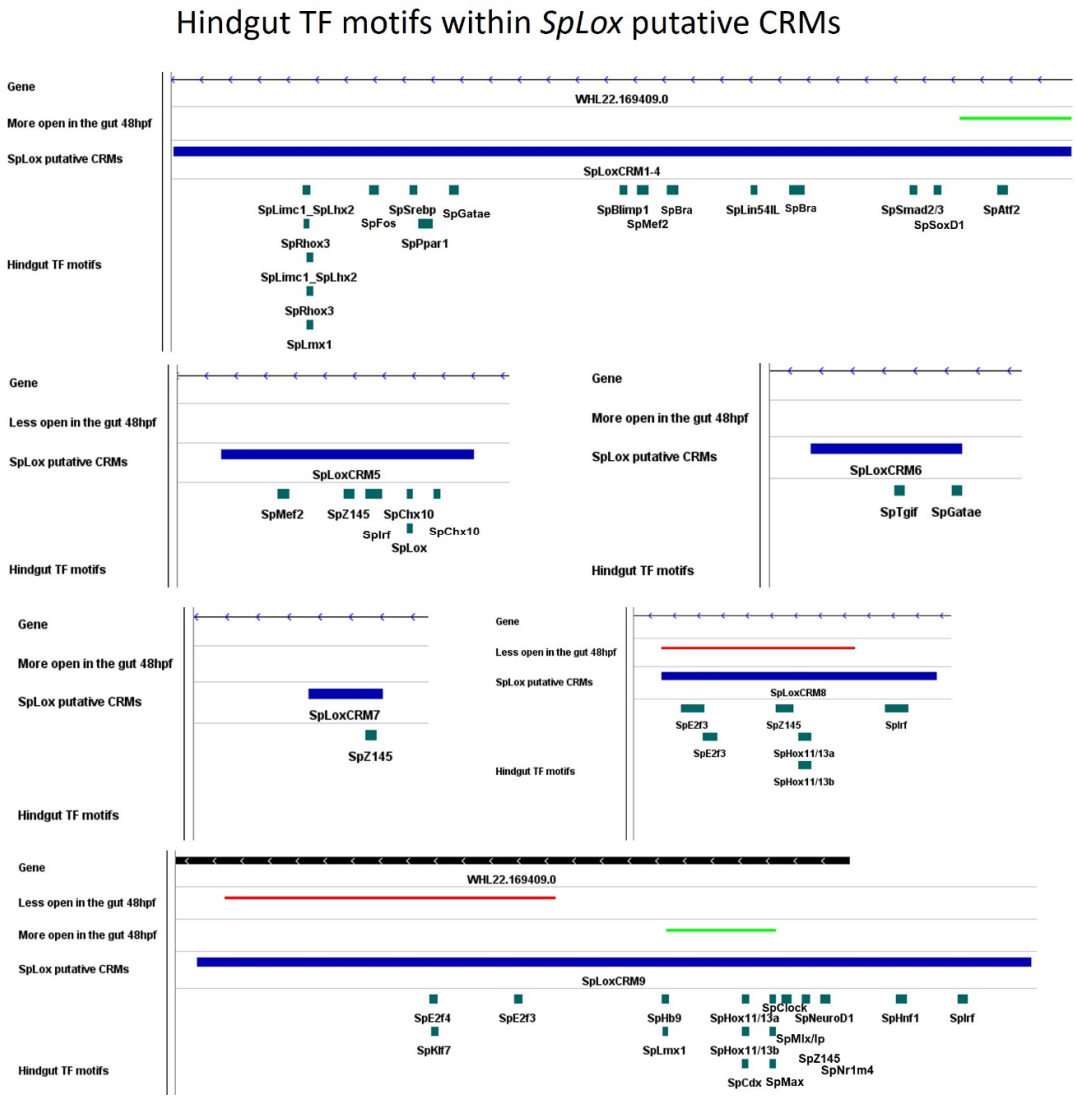


Figure 4.7 Potential transcription factor binding sites within putative *SpLox* CRMs.

Similar situation can be seen with the forkhead transcription factors that recognize similar DNA stretches in the CRMs around *SpCdx*, since SpFoxA, SpFoxD, SpFoxO and SpFoxI proteins can bind either to overlapping or same sequences according to the PWMs of homologous vertebrate TFs (Figure 4.8). There is only one locus in the *SpCdx* putative CRMs that is relatively more open

in the gut and it is located within SpCdxCRM1. As was stated before, this CRM and the sequence of the differentially open region within this CRM is conserved among the two sea urchin species (Figures 3.8 and 3.9), suggesting its conservation as a cis-regulatory module. Within this region SpBlimp1, SpAtf2 and SpCp2 can bind (Figure 4.8). Other regions of this CRM contain SpHox11/13b and SpCdx binding sites (Figure 4.8) suggesting a direct nature of the *SpCdx* self-regulatory loop (Annunziata & Arnone 2014; Annunziata et al. 2014). It is important to note that no SpLox motifs were found within the *SpCdx* putative CRMs pointing towards indirect nature of the known activation of *SpCdx* by SpLox (Cole et al. 2009). Same is the situation with SpBra, which is known to activate *SpCdx*, yet there are no SpBra binding sites predicted within the current list of putative CRMs (Figure 4.8). All the other known inputs such as SpFoxA and SpHox11/13b (Annunziata & Arnone 2014; Annunziata et al. 2014) into *SpCdx* could be direct as mentioned earlier (Figure 4.8).

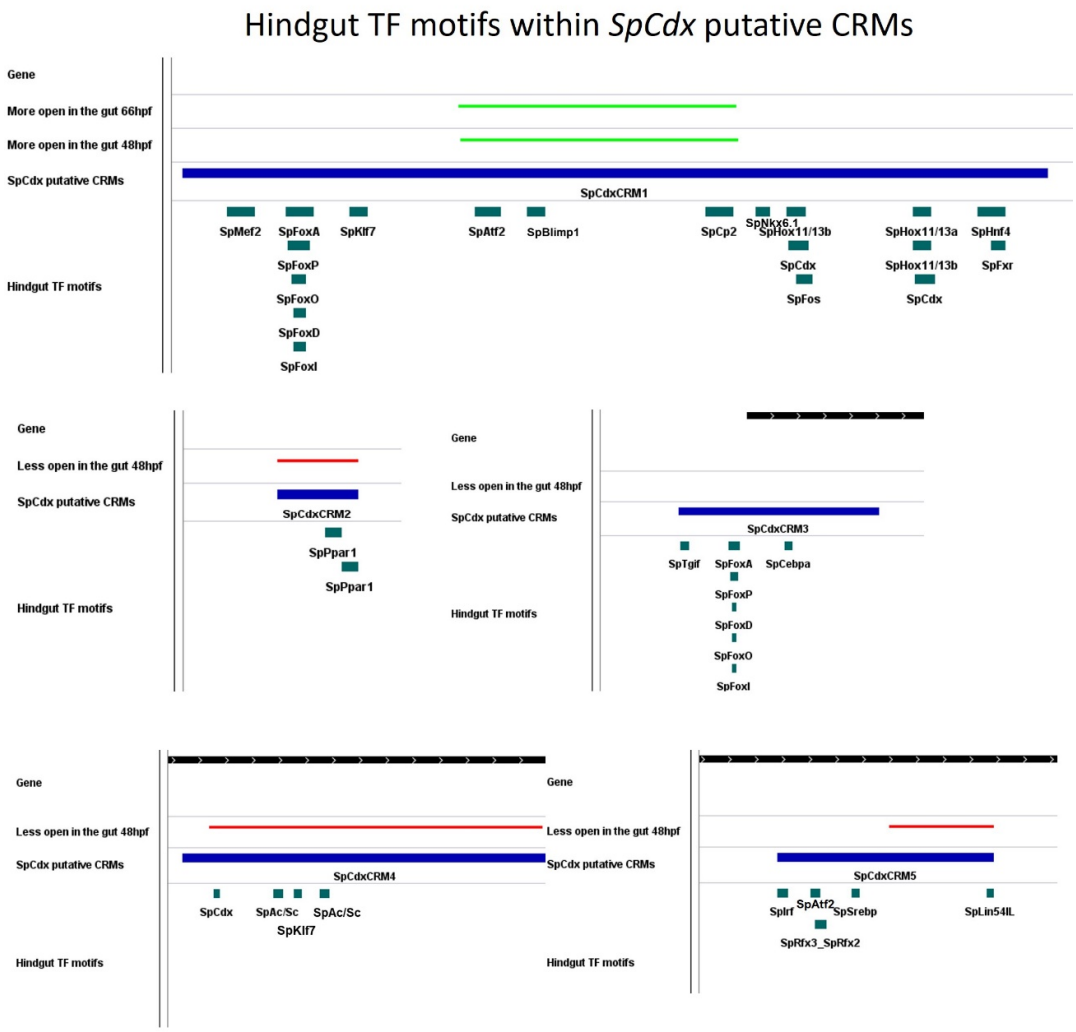


Figure 4.8 Potential transcription factor binding sites within putative *SpCdx* CRMs.

Functional assessment of the putative CRMs identified is described in the following chapter.

In addition to identifying transcription factors expressed in relevant domains, the interactions between all these transcription factors in terms of TF-CRM binding can also be predicted by assessing whether their vertebrate homologs (for which there are motif PWMs) have recognizable binding sites within the ATAC-peaks,

representing open chromatin, at 72 hpf. This analysis was performed and three GRNs, each for a hindgut cluster, were drafted (Figure 4.9).

The GRNs were drafted in order to assess the upstream control of the ParaHox genes. However, the resulting network topologies show that the majority of TFs that are capable of controlling expression of *SpLox* or *SpCdx* are also controlled by them (Figure 4.9). In total, 539 interactions were predicted: 143 for the pyloric sphincter, 239 for the intestine and 157 for the anus. Considering the approach used to predict these interactions, every interaction demonstrated in the GRNs could be direct, since putative cis-regulatory modules as well as transcription factors, capable of binding them, were identified for each node. The nodes shown at the top of the GRNs were identified to have inputs on other nodes of the GRN but no inputs on them (Figure 4.9). In case, of SpGatae this could indeed be true since its expression starts early and it is possible that, by 72 hours post fertilization, none of its targets or other genes in the GRNs involving the ParaHox genes can act as effectors of *SpGatae*. Nodes at the bottom of the GRNs constitute TFs that have binding sites in the merged putative CRMs of ParaHox genes, but these binding sites are inaccessible at 72 hpf (Figures 4.7, 4.8 and 3.8). The resulting GRNs recapitulate the published transcription factor nodes of the ParaHox GRNs in the hindgut at 72 hpf (Annunziata & Arnone 2014; Annunziata et al. 2014) and identify multiple new potential nodes and interactions at this developmental stage (Figure 4.9).

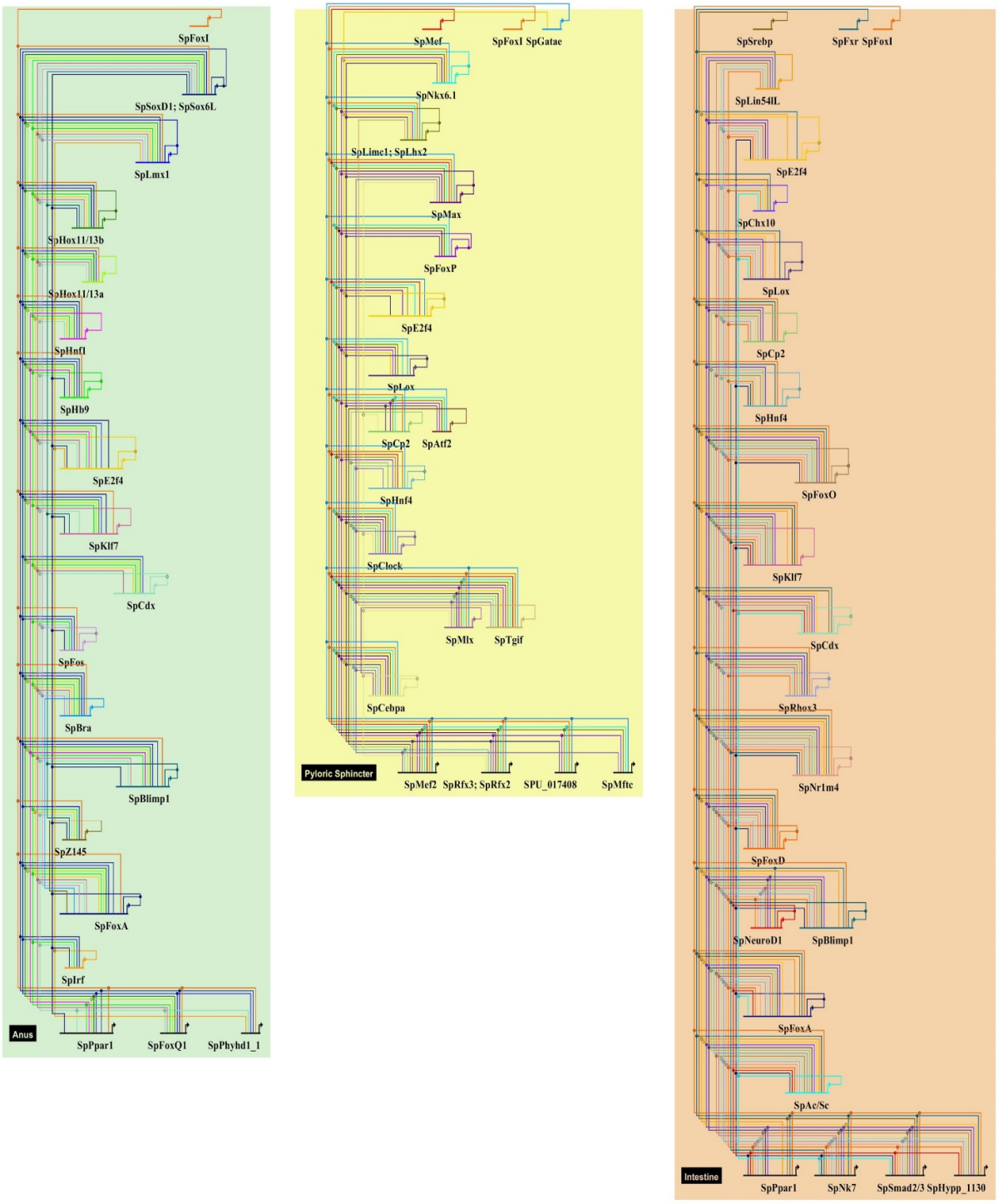


Figure 4.9 Draft GRNs upstream of the ParaHox genes for *S. purpuratus* pyloric sphincter, intestine and anus at 72 hpf.

4.3 Discussion

Combining chromatin accessibility data with transcriptomics allows drafting of the GRNs that recapitulate known topologies of these networks as well as predict new ones. For *S. purpuratus* such a blueprint of agents of control was drawn from the 72hpf pluteus ATAC-seq and scRNA-seq data. The scRNA-seq datasets give unprecedented resolution of gene expression analysis in the context of the molecular make up of the embryo. This resolution allows identification of tissue types and their transcriptomic profiles, allowing to narrow down motif matching software predictions, thus allowing to build GRN drafts of these tissue types. However, even then the number of potential interactions is likely to be overestimated (Peter & Davidson 2015). The GRNs presented in this chapter reflect the state of the hindgut regulation at the 72 hpf pluteus stage when the majority of cell types and organs of the sea urchin larva are already built. In order to be able to draw the temporal GRN dynamics, high resolution transcriptomic data is necessary from earlier stages of development in addition to the generated ATAC-seq data sets. The same applies to the other species of interest for evolutionary comparisons, since the high number of potential TF binding sites in the ParaHox CRMs discussed in this chapter means gross overestimation and also requires time and tissue specific transcriptomic datasets, that are not currently available. Problems with genome assemblies and gene annotations also pose a difficulty in such analyses. In light of this, generation of the scRNA-seq datasets for the *S. purpuratus* 48 hpf gastrula and the *P. lividus* 40 hpf pluteus is currently under way to start addressing GRN topologies in terms of temporal dynamics and evolution. These datasets are being produced as part of

collaboration with Periklis Paganos and Dr Jacob Musser. The new genome assembly for *S. purpuratus* has been produced and the new *P. miniata* genome sequencing is underway, with the *P. lividus* and the new *S. purpuratus* genome annotations in progress. These are undertaken by the sequencing consortia, which the lab of Dr Arnone is part of. These new genomes and their annotations will help improve the putative CRM predictions and also facilitate interspecies comparisons.

The described cis-regulatory modules and transcription factors are *in silico* predictions from NGS libraries. Lack of intrinsic means to confirm these, highlights the need for *in vivo* validations, through reporter gene constructs and gene perturbations that need to be performed.

CHAPTER 5

IN VIVO VALIDATIONS OF *IN SILICO* PREDICTIONS

This chapter pertains to experimental *in vivo* validations of the bioinformatical predictions of the ParaHox CRMs and their associated TFs.

5.1 Introduction

In the previous chapters an *in silico* approach to the GRN drafting was described. The ATAC-seq and scRNA-seq data allow predictions of cis-regulatory modules and their associated transcription factors (see chapters 3 and 4). However, these are simply predictions, which, in addition, do not give any information on the nature of the interactions between the GRN nodes. These interactions can be both positive, transcription factors bind enhancers to turn on or increase gene expression, and negative, in case of transcription factors binding silencers. Therefore, the activity of the putative CRMs needs to be shown *in vivo*, along with the transcription factor action validations.

Sea urchin embryos have a well developed methods tool-kit for such validations. Linear exogenous DNA injected into a sea urchin zygote becomes incorporated into the nuclear DNA of the developing sea urchin in a mosaic fashion (Franks et al. 1988). This mosaicism implies that not all the cells would have this exogenous DNA in their genomic DNA. However, such transgenesis allows efficient DNA reporter construct use for CRM validation (Arnone et al. 2004). The putative cis-regulatory elements would be combined with a basal promoter (such as SpEndo16 or SpGatae basal promoters) and a reporter gene such as genes coding for chloramphenicol acetyltransferase (CAT), luciferase (luc), β-galactosidase (lacZ) and green fluorescent protein (GFP) (Arnone et al. 2004). Use of such reporter constructs has lead to characterization of the cis-regulatory elements discussed in section 3.2.4.

In 2010 Dr Jongmin Nam and colleagues have developed a high throughput technique to test the activity of multiple putative cis-regulatory elements at once (Nam et al. 2010). They have synthesized sets of reporter tags, activity of which can be assessed via qPCR, using specific qPCR primers for each tag used. In other words, if three different CRMs combined with three different respective Tags are injected into a zygote and incorporated thus into the genomic DNA. Then the reporter gene in these tags is expressed if they are combined to an active CRM and each of the CRM-Tag constructs can be recognized using quantitative PCR due to the specific primers mentioned used for the qPCR reaction. Since 129 qPCR tags were synthesized by Dr Nam and colleagues, it became possible to test activity of 129 CRMs at once, and due to the high number of different CRMs within the same amount of injected DNA, the need to amplify every CRM-Tag construct became evident. Dr Nam with colleagues accounted for this in their Tag design allowing pre-qPCR amplification of every tag with universal primers that can amplify all of the 129 tags equally (Nam et al. 2010). Quantitative measurement of CRM-Tag constructs incorporated into the genomic DNA and expressed in the cDNA from the same embryos allows assessment of relative CRM activity.The combined CRM-Tag construct consists of the CRM, the basal promoter of SpGatae, the GFP open reading frame for the CRM activity visualization, the qPCR-identifiable Tag and the poly-A signal (Nam et al. 2010) (Figure 2.1).

Due to the high-throughput and quantitative capabilities this method was employed to test the ParaHox CRMs described earlier. Thus, the following sections of this chapter will describe the use of this 129 tag system to test the

putative *SpLox* and *SpCdx* CRMs identified during this thesis in order to functionally validate them.

5.2 Results

5.2.1 SpFoxA *CRM testing confirms known elements and increases resolution*

Prior to using the 129-tag method to validate the predicted CRMs for the ParaHox genes the method of prediction of putative CRMs from the ATAC-seq data (described in 2.5, 3.2.3, 3.2.6) had to be validated. In order to achieve this, *SpFoxA* CRMs from the ATAC-seq data were predicted. Out of the seven predicted CRMs five (SpFoxA_F_1, SpFoxA_I_1, SpFoxA_J_1, SpFoxA_K_1 and SpFoxA_K_2) overlapped with the known FoxA CRMs (Figure 5.1 A) that were published by de-Leon and Davidson in 2010. Putative CRMs SpFoxA_F_1, SpFoxA_FtoI_1, SpFoxA_J_1 and SpFoxA_K_1 were found to be differentially more open in the gut at 48 hpf compared to the whole embryo. Although SpFoxA_K_2 is less open in the gut at that time, it is more open at 66 hpf in the gut tissue dataset compared to the whole embryo (Figure 5.1), suggesting that this locus could play a role in the control of *SpFoxA* after the gastrula stage. Each of the putative seven *SpFoxA* CRMs were fused with a 129-tag reporter system Tag to visualize and quantify expression of these constructs to determine whether and when these CRMs are active and, if they are, the relative extent of this activity. First, a complete pool (all seven CRM-Tag constructs together) was injected in *S. purpuratus* zygotes in three biological replicates. The pools consistently reconstituted *SpFoxA* expression in the blastula endoderm and oral ectoderm (Figure 5.2 A and B) with less than 3% of the injected embryos showing

ectopic expression (mostly in the PMCs and misdeveloped embryos), as well as in the whole of the gut at the gastrula stage (Figure 5.2 C). The expression patterns of the pool of the *SpFoxA* putative CRMs was also concordant with the spatial expression driven by the FIJ CRM concatenate described in the de-Leon paper, which also recapitulated correct *SpFoxA* expression at 24 hpf in the endoderm and in the ectoderm (de-Leon & Davidson 2010).

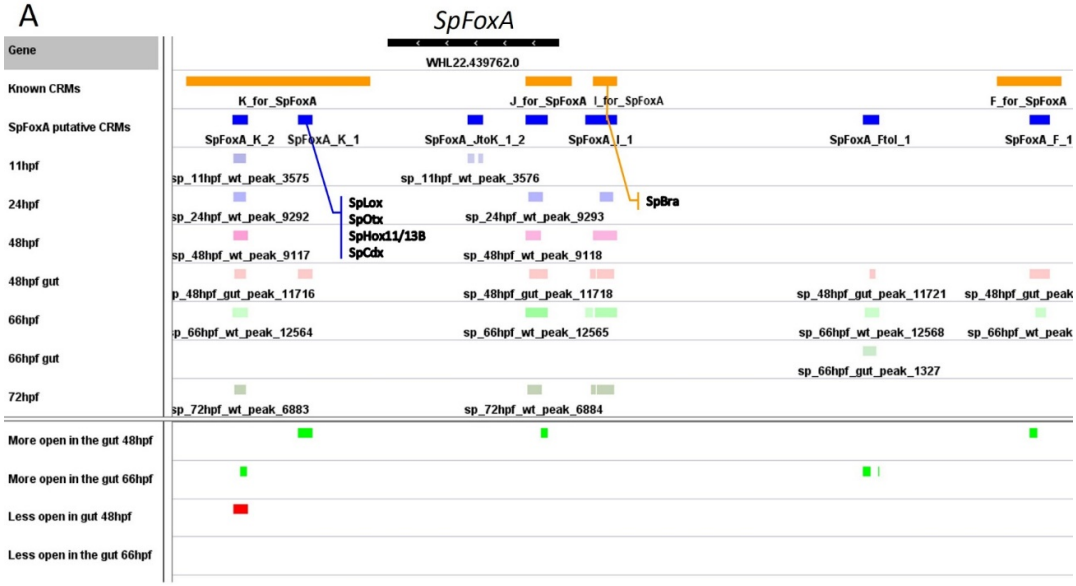
Scoring the embryos injected with the *SpFoxA* CRMs suggests batch effects, since two of the three batches showed less than 20% of the embryos showing the correct expression (Figure 5.1 B). Scoring of the SpFoxA CRM pool injected embryos, performed as part of the method validation is also concordant with the scoring done for the FIJ concatenate (de-Leon & Davidson 2010) with the majority of the expression being in the endoderm, while the oral ectoderm expression was visible in fewer embryos (Figure 5.1 B).

The quantitative expression assessment showed that two of the putative CRMs are most active at 24 hpf: SpFoxA_I_1 containing an SpBra binding site and SpFoxA_K_1 containing binding sites for SpOtx, SpHox11/13b and the ParaHox proteins SpLox and SpCdx (Figure 5.1 A). At 24 hpf, of course, the ParaHox genes do not have any effect on the expression of the constructs, since these genes are not yet transcribed, however SpLox could have a function at 48 hpf. SpFoxA_I_1 results support results from the paper in the context of the FIJ construct. However, region K containing SpFoxA_K_1 was not described in detail in the publication. Here, however, using the ATAC-seq predictions and the 129-tag system we show that it alone is enough to drive *SpFoxA* endodermal

expression as supported by it being the most active CRM of the pool and giving almost exclusively an endodermal expression profile (Figure 5.1 B and C; Figure 5.2 D). The high percentage of embryos expressing it could be due to the fact that all the exogenous DNA injected is contributed only by SpFoxA_K_1 and carrier DNA, without the inactive *SpFoxA* CRMs that can "dilute" the signal in the pool of embryos. The found SpHox11/13B and SpOtx transcription factor binding sites, mutation of which was shown to significantly change expression of the F-K concatenates compared to the wild type (de-Leon & Davidson 2010), present these two TFs as a likely candidates for early *SpFoxA* regulation in the endoderm through SpFoxA_K_1 CRM.

Here, therefore, we increase resolution of the *SpFoxA* regulatory regions by specifying, which loci of the ones previously identified give detectable expression. This supports the use of the putative CRM and TF prediction methods as well as utilization of the 129-tag system constructs to validate the ParaHox CRMs.

A



B

Scoring of embryos injected with *SpFoxA* CRM-Tag constructs

Batch	Embryos examined	Endodermal and Ectodermal	Endodermal	Ectodermal	Correct	Ectopic	Total
Pool 1	93	9 (9.7%)	16 (17.2%)	12 (12.9%)	37 (39.8%)	3 (3.2%)	40 (43.0%)
Pool 2	99	0 (0.0%)	11 (11.1%)	2 (2.0%)	13 (13.1%)	3 (3.0%)	16 (16.2%)
Pool 3	110	2 (1.8%)	12 (10.9%)	4 (3,6%)	18 (16.4%)	4 (3.6%)	22 (20.0%)
CRM K1	56	0 (0.0%)	24 (42.9%)	0 (0.0%)	24 (42.9%)	5 (8.9%)	29 (51.8%)

C

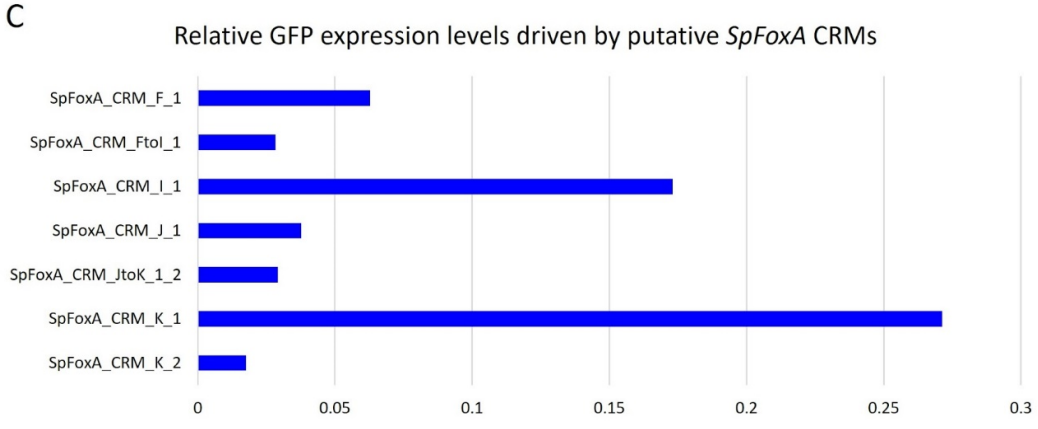


Figure 5.1 *SpFoxA* putative CRM testing A. Putative *SpFoxA* CRMs in relation to *SpFoxA* gene, published CRMs, ATAC-seq peaks and differentially accessible loci within. Binding sites for TFs are labelled with blue and orange connector lines. B. Scoring table for embryos injected with

SpFoxA CRM-Tag constructs at 24 hpf. C. Relative GFP expression levels driven by the SpFoxA CRMs at 24 hpf.

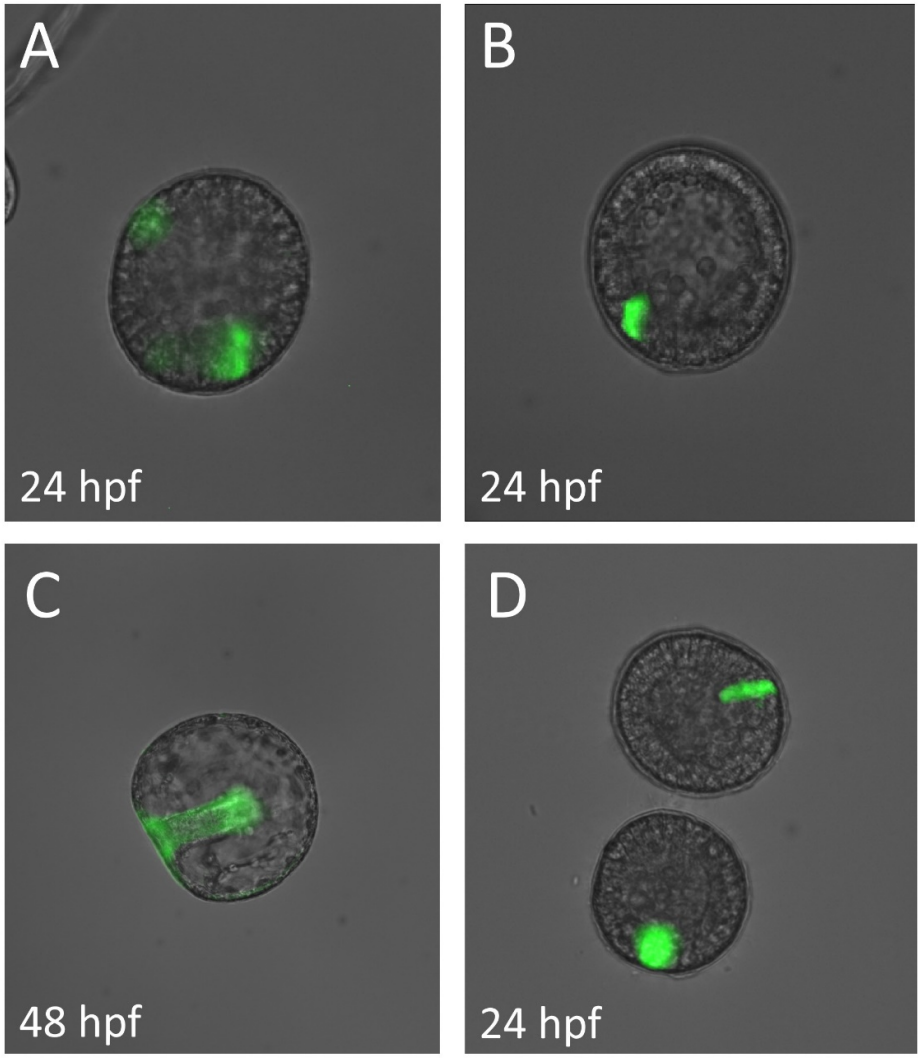


Figure 5.2 GFP expression driven by the *SpFoxA* CRMs. A, B. GFP expression driven by *SpFoxA* CRM pool at blastula stage. C. GFP expression driven by the *SpFoxA* CRM pool at the gastrula stage. D. GFP expression driven only by *SpFoxA*_K_1 at the blastula stage. Green is GFP.

5.2.2 SpLox *putative CRMs validation: verification of direct SpHox11/13B input*

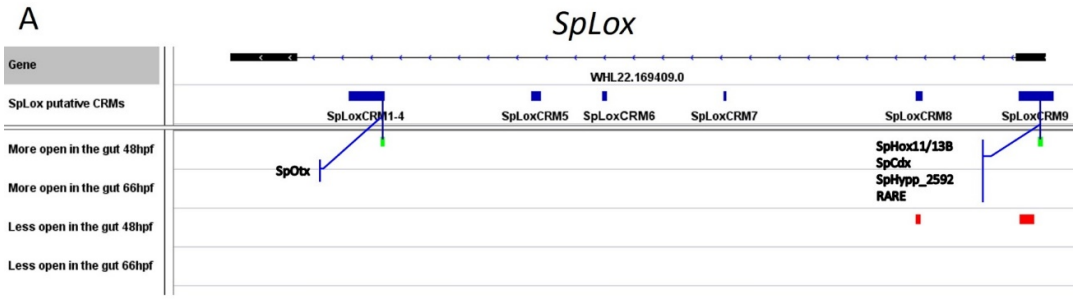
Following the support of the approach used by *SpFoxA* CRM testing, putative *SpLox* CRMs have been analyzed in the similar manner. A pool of all the six

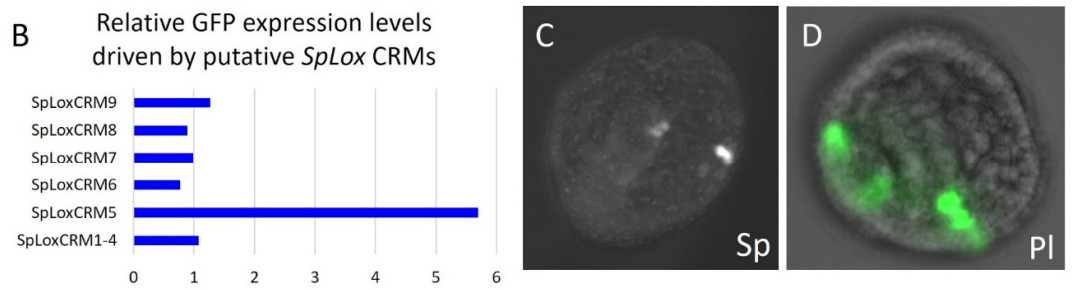
SpLox CRMs was injected into *S. purpuratus* zygotes and its expression at 48 hpf, around the time when *SpLox* expression peaks (Arnone et al. 2006), was assessed. Two of them are more open in the gut compared to the whole embryo: SpLoxCRM1-4 and SpLoxCRM9 (Figure 5.3 A). The injected CRM pools reconstitute the *SpLox* expression pattern at this time point (Arnone et al. 2006), with the CRM expression visible in the middle of the hindgut and in the ectoderm (Figure 5.3 C). The ectodermal expression of *SpLox* in the lateral neuronal ganglions has also been shown and described (Perillo et al. 2018). Quantitative expression analysis on these showed that SpLoxCRM5 is the most highly expressed (Figure 5.3 B), however this is likely to be a replicate issue (replicate 1 showed relative GFP expression adjusted to incorporated DNA of 2.2, while replicate 2 showed expression of 9.2). The next most active *SpLox* putative CRM is SpLoxCRM9. In general, expression levels of *SpLox* CRMs were not found as different as for *SpFoxA* CRMs. SpLoxCRM9 was chosen for in-depth assessment due to its high relative expression, existence of both a more open chromatin region in the gut dataset and less open chromatin regions compared to the whole embryo (Figure 5.3 A), and the gut specific binding sites identified within the open region (Figure 5.3 and Figure 4.7). The putative SpLoxCRM9 was injected separately into the sea urchin fertilized egg and was shown to be able to drive *SpLox*-like expression pattern in the 48 hpf gastrula: in the hindgut (near the forming pyloric sphincter) as well as in the lateral ganglion neurons (Figure 5.3 E, F and G). This suggests importance of this CRM as the regulator of *SpLox* expression in both germ layers.

A putative transcription factor affecting this region was also tested by co-injecting SpHox11/13B morpholino (MO) that prevents translation of this transcription factor with SpLoxCRM9. This resulted in a drastic decrease of GFP expression driven by this CRM in the endoderm (15.4% of all injected embryos with just SpLoxCRM9 vs 1.4 % of all injected embryos with both SpLoxCRM9 and SpHox11/13b MO) (Figure 5.3 E), and in general a 31% decrease in total number of embryos with expression (p-value= 3.339225663898966e-20). This points to the direct activation of *SpLox* by SpHox11/13b via SpLoxCRM9 at the 48 hpf gastrula. One of the binding sites for this transcription factor is found in the more open region of the chromatin located within this CRM. This relatively more accessible region was separately amplified and combined with the reporter tag, and it shows expression in the hindgut around the forming pyloric sphincter (Figure 5.3 H). Suggesting, indeed, that this open region is responsible for control of *SpLox* transcription in the gut.

Sequence conservation discussed in section 3.2.6 points to conservation of *Lox* gene cis-regulatory elements in the two sea urchin species: *S. purpuratus* and *P. lividus*. Injection of the pool of six *S. purpuratus Lox* CRMs into the fertilized eggs of *P. lividus* gives the same expression pattern at 24 hpf gastrula in the Mediterranean sea urchin species as in the Pacific species 48 hpf gastrula, which is concordant with *Lox* expression. The GFP expression, again, is visible in the neurons and hindgut of the *P. lividus* gastrulae (Figure 5.3 D). This not only points to CRM sequence conservation between these species, but also suggests the existence of similar transcription factor machinery in the same regions capable of driving gene expression through these similar CRMs.

These results show that putative SpLoxCRM9 is functional, capable of controlling *SpLox* in two germ layers, that it is driven by SpHox11/13b, and that it has a gut active locus. High similarity of most of its sequence with PlLoxCRM3 suggests that the *P. lividus* CRM has the same function, subject to confirmation.





E

Scoring of embryos injected with *SpLox*CRM9-Tag construct

	Total embryos	Total fluorescent	Ectoderm only	Endoderm only	Ectoderm & Endoderm
SpLoxCRM9	175	67	26	27	14
Percentage of total embryos	100%	38.3%	14.9%	15.4%	8%
Percentage of total fluorescent		100%	38.8%	40.3%	20.9%
SpLoxCRM9 + SpHox11/13B MO	212	16	13	3	0
Percentage of total embryos	100%	7.5%	6.1%	1.4%	0%
Percentage of total fluorescent		100%	81.25%	18.75%	0%

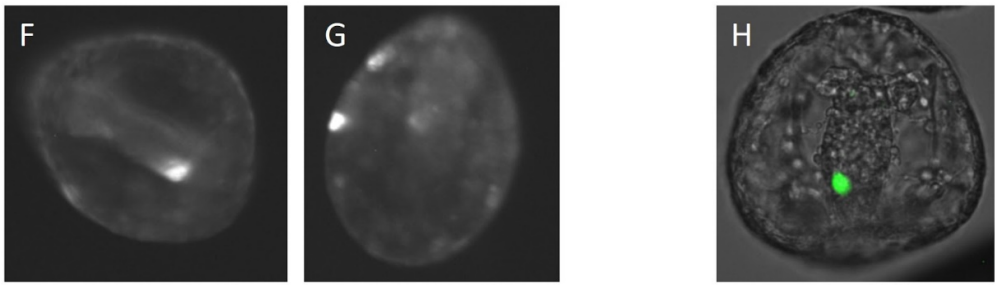


Figure 5.3 *SpLox* CRMs validations. A. Putative *SpLox* CRMs in relation to *SpLox* gene and differentially accessible loci within. Select binding sites for TFs are labelled with blue connector lines. B. Relative GFP expression levels driven by *SpLox* CRMs at 48 hpf. C. GFP expression

driven by *SpLox* CRM pool at 48 hpf (white sites). D. GFP expression driven by *SpLox* CRM pool in *P. lividus* at 24 hpf (green sites). E. Scoring table for embryos injected with SpLoxCRM9 constructs. F, G. GFP expression driven only by whole SpLoxCRM9 at 48 hpf (white spots). H. GFP expression driven by only a sub-region of SpLoxCRM9 more open in the gut compared to the whole embryo at 48 hpf (green spot).

5.2.3 SpCdx *putative CRMs validation shows ectopic activity of SpCdxCRM1*

The putative CRMs for the second *S. purpuratus* gut ParaHox gene, *SpCdx,* were also tested around the time when its transcription reaches its peak, in this case at the 66hpf prism stage. Three of five of the SpCdxCRMs have regions less accessible in the gut samples at 48 hpf compared to the whole embryo, however one CRM, SpCdxCRM1, which has sequence similarities with *P. lividus* PlCdxCRM1 and PlCdxCRM2 and a shorter shared sequence with *P. miniata* PmCdxCRM7 (see section 3.2.6), has a region that is relatively more accessible in the gut at 48 hpf and at 66 hpf (Figure 5.4 A). First, a pool of all five putative SpCdxCRMs was injected in the zygotes. Expression at 66 hpf does indeed show expression concordant with endogenous *SpCdx* expression in the posterior-most regions of the hindgut (Cole et al. 2009): the intestine and the anus (Figure 5.4 B and D; Figure 4.3), however, a considerable percentage of the injected embryos showed ectopic expression in the forming pyloric sphincter (Figure 5.4 B and E).

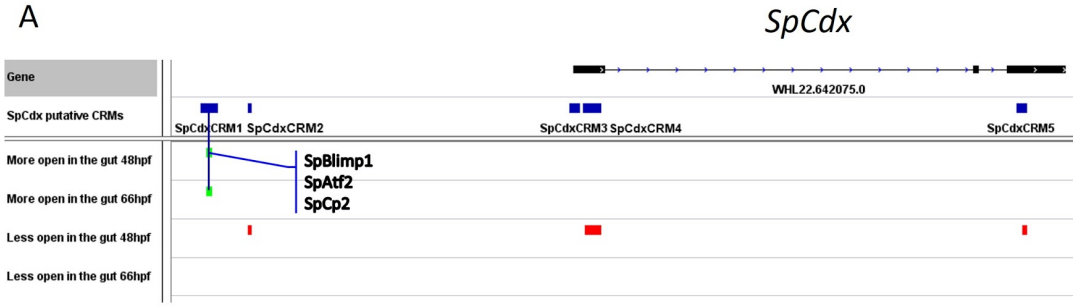
SpCdxCRM1, as mentioned, has relatively more accessible chromatin regions in the gut enriched datasets, contains binding sites for TFs such as SpCdx, SpFoxA, SpNkx6.1, SpAtf2, SpBlimp1 and SpCp2 (Figure 5.4 A and Figure 4.8) and is one of the CRMs driving GFP expression (Figure 5.4 C). This CRM was separately injected into the fertilized *S. purpuratus* egg and the GFP expression pattern that it drives is ectopic to endogenous *SpCdx* expression (Cole et al. 2009), since

GFP expression driven by this CRM is detectable at the top of the stomach near the cardiac sphincter and at the pyloric sphincter (Figure 5.4 B and G). This expression is similar to the sphincter expression obtained with the whole pool of SpCdxCRMs (Figure 5.4 E). Such an expression pattern could be due to SpNkx6.1 which is expressed in these regions, however, if this CRM targets *SpCdx*, it is likely that its input on *SpCdx* is repressed by the TFs bound within other SpCdxCRMs, which in case of injection of a single CRM are missing from the regions where the injected construct is incorporated. Still the interspecies sequence similarity of this CRM suggests that it is conserved, indicating its importance for regulation of *Cdx.*

This suggests that other SpCdxCRMs contribute to the *SpCdx* expression in the posterior hindgut, which is also supported by the qPCR data, since SpCdxCRM3 and SpCdxCRM4 are relatively more active in driving GFP expression (Figure 5.4 C).

The *SpCdx* CRM transgenesis shows that the SpCdxCRM1, with the region open in the gut, is not responsible for *SpCdx* expression in the posterior hindgut, and that the endogenous expression is attributable to other SpCdxCRMs.

A

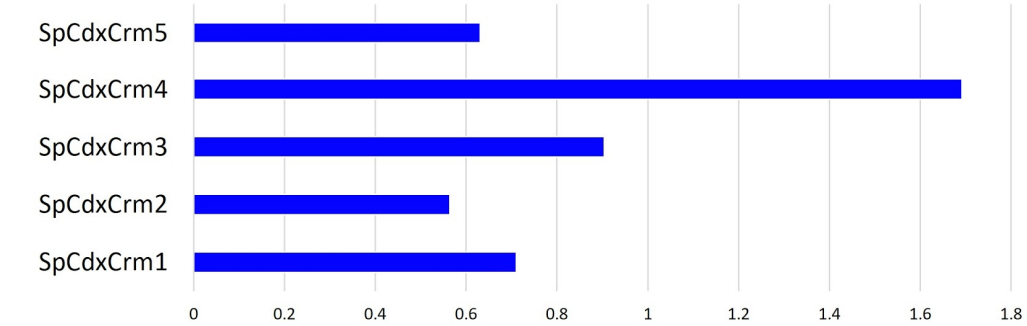


B

Scoring of embryos injected with *SpCdx* CRM-Tag constructs

Batch	Embryos examined	Correct	Ectopic	Total
Pool 1	53	10 (18.9%)	6 (11.3%)	16 (30.2%)
Pool 2	57	5 (8.8%)	4 (7%)	9 (15.7%)
SpCdxCRM1	97	0 (0%)	30 (30.9%)	30 (30.9%)

C



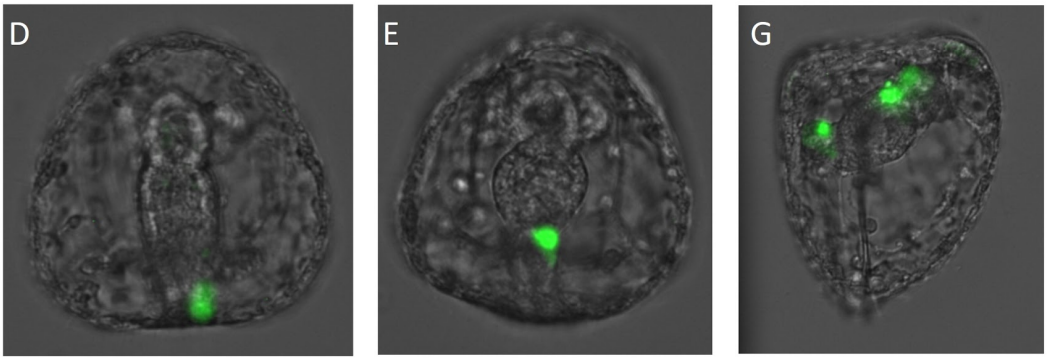


Figure 5.4 *SpCdx* CRM validations. A. Putative *SpCdx* CRMs in relation to *SpCdx* gene and differentially accessible loci within. Select binding sites for TFs are labelled with blue connector lines. B. Scoring table for embryos injected with *SpCdx* CRM-Tag constructs at 66 hpf. C. Relative

GFP expression levels driven by the *SpCdx* CRMs at 66 hpf. D. GFP expression driven by the pool of SpCdxCRMs at 66 hpf corresponding to *SpCdx* expression pattern at 66 hpf. E. GFP expression driven by the pool of SpCdxCRMs at 66 hpf ectopic to *SpCdx* expression pattern at 66 hpf. G. GFP expression driven only by SpCdxCRM1 at 66 hpf. Green is GFP.

5.3 Discussion

The *in vivo* validations performed so far highlight the robustness of the *in silico* approach. Testing of the putative *SpFoxA* CRMs in comparison with the published information on the cis regulation of this gene shows, along with other known CRMs discussed in sections 3.2.4 and 4.2.1, that, indeed, chromatin accessibility essays such as ATAC-seq are valuable tools in CRM recognition. *SpLox* CRM validations allowed to confirm the predicted direct positive effect of SpHox11/13b on expression of *SpLox* through binding of SpLoxCRM9. On the other hand, *SpCdx* CRM validations so far failed to point out the exact CRMs responsible for the observed *SpCdx* expression pattern, however, this work is in progress. Interestingly, the SpCdxCRM1, which has open chromatin in the gut enriched ATAC-seq samples compared to the whole embryo, does drive expression of GFP in the gut, although in the "wrong" regions compared to *SpCdx* expression.

It is worth noting that many CRMs and their transcription factors could only have quantitative effects (Wahl et al. 2009; Cui et al. 2017b; Peter & Davidson 2015) and as such they could be unable to drive GFP expression at a high level on their own or in a pool during transgenetic experiments, where they integrate in random locations in the genome (Franks et al. 1988). Concatenations of the CRMs will be used to assess these quantitative effects.

Overall, these results show the groundwork for building the gene regulatory network around the ParaHox genes and suggest that the approach used for doing it is valid. Highlighting, however, the need for more work to be done, which will be discussed in the final chapter of this thesis.

CHAPTER 6

DISCUSSION

This chapter will concern the general discussion of the methodologies used and the results obtained, as well as suggest the future direction of the research.

6.1 Combinatorial approach is effective at GRN drafting

This project concerns itself with a gene regulatory network upstream of the ParaHox genes in *S. purpuratus, P. lividus*, *P. miniata* and *B. lanceolatum.* It is possible to use an unbiased approach, developed in collaboration with Dr Lowe, to drafting of this GRN using various omics data sets, such as ATAC-seq, ChIP-seq, RNA-seq and scRNA-seq, without prior knowledge of any of its components except the two ParaHox genes that play the role in the gut development. The approach is somewhat incremental and allows building the complexity of the GRN step by step. First comes the identification of the nodes of interest (for this project these are the ParaHox genes *Lox* and *Cdx*), then CRMs are predicted from the ATAC-seq data or ChiP-seq data and their putative TFs are predicted using bioinformatical tools such as HOMER, which is followed by filtering of these PWM based transcription factor predictions by the time and tissue specific expression data obtained through differential RNA-seq or scRNAseq (for the *in silico* drafting 72 hpf scRNAseq data was used in the context of this thesis). After narrowing down the transcription factors that could bind putative ParaHox CRMs, the same ATAC-seq data can be used to predict whether genes coding these TFs have putative CRMs themselves, and then the ability of all the putative TFs targeting the ParaHox genes to target the CRMs near the genes coding these TFs can also be assessed. In other words, this method allows prediction of all inputs affecting the gene or genes of interest as well as their interactions between each other, making the *in silico* gene regulatory network drafting possible, given the availability of the data at a given timepoint for a given tissue. Comparing the predictions obtained through the bioinformatical means with the previously

published CRMs as well as the validations performed during the thesis show the predictory robustness of the approach.

The different types of approaches used and the variety of sequencing data obtained during this project, therefore, allows to look at a GRN from multiple perspectives. Using the bioinformatical and laboratory tools described, it is possible to identify multiple putative cis-regulatory regions, their target genes, as well as to predict the transcription factors that could bind these CRMs, thus giving the nodes of the GRN and suggesting the links between them. The *in vivo* validations allow identification of the nature of the links, whether they are positive (activating) or negative (inhibiting).

6.2 Issues with data and software

Of course, this method of uncovering network topologies is not perfect. The issues mainly include problems with the library preparations and the bioinformatical tools used. As discussed in chapters pertaining to the differential gene expression in the gut compared to the whole embryos and the differential accessibility of chromatin in the same conditions, the technique of extracting the gut samples from the echinoderm embryos is not perfect. This method involves the chemical treatment to deplete calcium and the mechanical separation of the cells surrounding the gut (see section 2.2), this physical separation means that the cells located at the extreme ends of the linear gut can be lost. The differential analysis results suggest that this, indeed, happened, with the genes expressed in that posterior-most gut lost in all of the gut samples. However, the genes specific to the coelomic pouches suggest that, conversely, these tissues did not

get separated from the gut tissue. Due to the small size of the extracted gut tissue, the assessment of its composition during the extraction is difficult. This leads to incomplete libraries for ATAC-seq and RNA-seq. Another experimental issue is the fact that extraction of both the genomic DNA and RNA from the embryos using AllPrep DNA/RNA Micro kit (Qiagen) seems to be dependent on the number of embryos lysed, which could have lead to inconsistencies between the replicates, observed in the qPCR validations described. In addition, animals are not equally responsive to transgenesis, as shown, for instance, by *SpFoxA* CRM injections, which also could lead to some differences in replicates.

Majority of issues relevant to the results obtained during this thesis lie in the bioinformatical methods used. For instance, there are no sea urchin, sea star or amphioxus specific transcription factor motif databases, to the author's knowledge. JASPAR for instance has motifs for vertebrates, insects, plants, fungi, urochordates (one PWM) and nematodes (Khan et al. 2018). The closest available database is for vertebrates, and, thus, it was used for motif matching to predict transcription factor binding sites. This could lead to misidentification of the binding sites in the species of interest, in case their transcription factors do not recognize the same sequences as their vertebrate counterparts. Further, this means that the homologs for these vertebrate transcription factors need to be found for each species. This depends on the availability of the transcriptomic and proteomic data for a given species, which could result in misidentification of the homologs. However, transcription factors are conserved and the validations indicate that the vertebrate PWMs are valid for predictions in the sea urchin.

HOMER was also used for *de novo* motif predictions within relatively more accessible genomic regions in the gut compared to the whole embryo. This software was designed to discover transcription factor motifs *de novo* in ChIP-seq peaks. ChIP-seq peaks are specific to a particular transcription factor and should contain sequences for recognition by this transcription factor. ATAC-seq peaks however are more heterogeneous, so prediction is less accurate and some *de novo* motifs could be false positives. Identification of many motifs expected in the explored samples, again, highlights suitability of this tool for performing the analyses described.

HiC data sheds light on the three dimensional chromatin organization and identifies physical interactions between chromosomal loci, and shows this through interaction count matrices, but there are few software packages developed for testing the significance of these interactions (Carty et al. 2017; Mifsud et al. 2017; Heinz et al. 2010). Due to sparsity of the identified counts at high resolution, actual interactions could be considered insignificant. More replicates are required to explore the significance of the identified interactions between the ParaHox loci in the sea urchin.

6.3 Evolutionary comparisons

Evolution of gene regulatory networks can be driven through different components of these network: the nodes, described through genes, their products (proteins) and their associated cis-regulatory regions, as well as their interactions. This thesis attempts to perform comparisons between these components in four deuterostome species: two sea urchin species *S. purpuratus*

and *P. lividus*, a sea star species *P. miniata* and an amphioxus species *B. lanceolatum*. Homology of genes between these species makes such comparisons possible. In terms of genes, a single copy of each ParaHox gene exists in every species of interest. Organization of these genes on the chromosomes of these species suggests that an intact cluster is necessary to keep the same gene orientations: in *P. miniata* and *B. lanceolatum* the genes are oriented the same way, while in the sea urchins, where the cluster is broken up, the relative gene orientations are different. This notion, along with the lack of clustering itself, could have implications to enhancer sharing. It is possible that species that retain the intact cluster with the same orientations could use the same enhancer to drive genes in a spatial and temporal manner leading to collinearity. Ectopic expression of a single *SpCdx* CRM (SpCdxCRM1) could account for that since its parts are conserved among the three echinoderm species, and it is not driving expression in the *SpCdx* location, possibly due to the repression of this enhancer around the actual *SpCdx* gene, absent in the injected construct. This enhancer needs in-detail analysis in the context of the three species.

The conserved CRM sequences between *S. purpuratus* and *P. lividus* have the same functions in these closely related species since *S. purpuratus* CRMs can drive GFP expression in the same regions in both species, showing presence of necessary transcription factors to activate the CRMs. Sequence similarities between the two species indicate that the transcription factors capable of activating SpCdxCRMs in *P. lividus* also activate their own *PlCdx* CRMs. This

finding suggests that at a short evolutionary distance, as is between the two sea urchin species, the GRN components didn't change much.

However, TF binding sites driving CRMs can have high turnover rates (Moses et al. 2006) and in more distant species such as *B. lanceolatum* the same transcription factors can drive same genes, but through different CRMs and, therefore, the actual binding sites can be in a different location in every species. In attempt to gain insight into this, TF predictions were compared among the species. The results in the section 4.2.1 show that many predicted TFs bound are indeed the same, with the motif occurrence counts suggesting similarities between transcription factor repertoires of the two sea urchin species, and somewhat between all three echinoderms while amphioxus is more different. However, the motif counts in this case may not indicate which TFs are actually important and HOMER motif match gives too many results for any sensible comparisons, therefore, in order to assess GRN evolution between the deuterostomes of interest, this number needs to be filtered, which so far was only possible for the *S. purpuratus* 72 hpf pluteus stage through the generation of the high resolution single cell RNA-seq data.

6.4 Future outlook

6.4.1 More transcriptomic data required

The need for bioinformatical filtering of the predicted transcription factor binding sites, along with other considerations, orchestrates the future outlook. The lack of transcriptomic data to perform this filtering indicates the need to produce expression assay datasets for the species in question. The scRNA-seq data for

S. purpuratus is an invaluable source for transcription analysis. Current data concerns only the 72 hpf pluteus, however to assess the developmental trajectories and pin-point the genes expressed in the relevant tissues earlier in development scRNA-seq datasets for other embryonic stages are necessary. To this end, a scRNA-seq dataset for *S. purpuratus* gastrula at 48 hpf is under construction. To perform inter-species evolutionary comparisons single cell transcriptomes of *P. lividus* plutei are also being established, and in the long run, the same sets of transcriptomic data for *P. miniata* and *B. lanceolatum* from different developmental stages need to be generated. Otherwise, in case of the ParaHox genes the transcriptomes of cells expressing these genes can also be generated after fluorescence-activated cell sorting (FACS) (Herzenberg et al. 1976) using signals from *Lox* or *Cdx* antibodies. This will also allow filtering of the predicted TF repertoire, however due to roles of ParaHox genes in neurons, this filtering cannot be gut specific without identifying a gut marker for co-sorting in each species.

6.4.2 Single cell ATAC-seq

The resolution of uncovering of GRN topologies can also be improved through the use of single cell ATAC-seq (Cusanovich et al. 2018). This, combined with the single cell RNA-seq can give both chromatin accessibility and the associated transcription profiles from the same cell. Such approach would greatly increase specificity of forecasting of nodes and interactions within a GRN from a given tissue type, for instance, the intestine or the anus.

6.4.3 Improved genome assembly for P. miniata

Quality of a genome assembly and gene annotations were discussed to have an impact on identification of putative CRMs (sections 3.2.6 and 3.3). Therefore, an improved *P. miniata* genome assembly is necessary to fill the missing genomic information between *PmGsx* and *PmLox* to facilitate the assessment of open chromatin regions and correctly point out the cis-regulatory regions. This work is undertaken by a genome sequencing consortium of a number of labs, and the improved assembly should be available in the near future.

6.4.4 Retinoic acid control of ParaHox genes

The ParaHox genes in amphioxus are under retinoic acid control (Osborne et al. 2009), with retinoic function conserved in chordates (Kam et al. 2012; Beckett & Petkovich 1999). Motif scanning of putative ParaHox CRMs in *S. purpuratus* has identified retinoic acid response elements within (Figure 5.3). This, along with the work performed on the function of retinoic acid in sea urchin embryos by Dr Rosella Annunziata, suggests that retinoic acid controls expression of the ParaHox genes through the interaction of retinoic acid related transcription factors with putative ParaHox CRMs. Analysis of the chromatin accessibility assay from retinoic acid treated embryos is also an ongoing work. This analysis will allow identification and validation of the regions of DNA accountable for the response to retinoic acid in sea urchin development.

6.4.5 Synthetic enhancers, ATAC-seq and HiC

Transgenic experiments show that an active CRM is capable of driving GFP expression in a sea urchin embryo in the expected spatial pattern, which corresponds to the endogenous expression of the target gene, and in other cases in locations that are ectopic to the target gene locations. Gene expression, as discussed in the introduction, is dependent on chromatin looping by CRMs and transcription factors bringing gene promoters close transcription machinery. In the "wild type" embryos and the nuclei of their cells, this looping is, therefore, responsible for the activity of the CRM on the target gene. In the sea urchin embryo the linear exogenous expression vectors are incorporated randomly into the DNA, which implies that they do not have the same spatial constraints as the actual CRMs in the genome do. However, these exogenous constructs need to be accessible in the chromatin to allow the transcription of the GFP reporter and the transcription factors that are responsible for their activity still need to recruit the transcription machinery. The mechanisms of this are particularly interesting in the context of expression occuring at transcription factories. The protocols developed for ATAC-seq chromatin accessibility assessment in the sea urchin (Magri et al. in press) and the possibility of 4C/HiC experiments, make this experimental system suitable for studying of mechanisms of such transgenesis and of function of exogenously introduced CRMs. Such experiments can be performed with designing a synthetic CRM (to be able to identify it in the genome as exogenous) with functional binding sites for a transcription factor with a known expression pattern and then assessing the chromatin accessibility and the three dimensional organisation of the sites of integration.

6.4.6 CRISPR/Cas9 and mutagenesis

Results presented in chapter 5 show the use of morpholino to prevent translation of a predicted transcription factor shown to control this CRM, however, this, still, does not definitively confirm that this control is direct. In order to do this, TF binding sites within the CRMs will be mutated by PCR using primers to introduce a different sequence (PCR site directed mutagenesis) (Ling & Robinson 1997). This will remove the particular binding site from the CRM and validate direct recognition and control by the identified transcription factor. In addition, with the advent of CRISPR/Cas9 approaches that are possible to perform in the sea urchin (Oulhen & Wessel 2016), a perspective approach to study cis-regulatory regions would be to change a binding-site sequence “in place” in the actual CRMs in the genomes. If such a change affects the expression of the target gene it would also confirm the direct effect of the transcription factor recognizing that binding site on the target gene.

6.5 Conclusion

The aims and goals of this thesis are stated in the introduction chapter.The work done during this thesis has achieved the goals set, contributing to the attaining the aim of elucidating the mechanisms that control the expression of *Lox* and *Cdx* genes in four deuterostome species: *S. purpuratus, P. lividus, P. miniata* and *B. lanceolatum* and their evolution. Given the new genomic data, the chromosomal arrangement of the ParaHox genes was finally confirmed for the sea urchin species to not be arranged in a cluster, unlike the organization of these genes in the sea star and amphioxus, which is likely to be ancestral (Ikuta et al. 2013).

Expression of these genes with the intact cluster suggests the spatial and temporal collinearity, while in the sea urchins, that don't have this cluster, the temporal expression is reversed, while the spatial collinearity is kept. The three dimensional organization of the Hox genes in their cluster in the genome was suggested to be important for regulation of this collinearity by opening and closing the chromatin along the cluster (Gaunt 2015). However, most open chromatin regions around the ParaHox genes persist throughout the developmental stages in all species suggesting a different way of controlling the collinearity of the ParaHox genes (Ikuta et al. 2013).

In addition, enhancer sharing was also proposed to be important for the collinearity of the ParaHox gene expression (Graham et al. 1989). Within the species where the cluster exists this is possible due to gene proximity on the scaffold. In species with a dispersed cluster this is only possible if the genes are close to each other in three-dimensional organisation of the chromatin. Assessing if this is the case for the sea urchin species was the second goal of this thesis. The analysis performed (section 3.2.2) suggests that there is no significant physical contact between the ParaHox loci in *S. purpuratus*, therefore enhancer sharing in the sea urchin is likely impossible. This could explain the absence of the correct temporal collinearity in the sea urchin species, therefore, analysis of which enhancers, targeting the ParaHox genes in the species where these genes are in the cluster, could be shared between different genes is necessary.

Therefore, identification of putative cis-regulatory regions was defined as the third goal and transcription factor inputs on them as the fourth. This work resulted in

the identification of the putative CRMs in the four species through the use of open chromatin ATAC-seq data. In addition, the potential transcription factor binding sites were also identified within these. However, due to the vast number of TF predictions, more information, such as time and tissue specific expression data is necessary to narrow down the number of transcription factors potentially binding to the CRMs. The method for doing it was developed and explored within one of the sea urchin species *Strongylocentrotus purpuratus*, allowing to draft a GRN at a certain time-point (the GRN draft of the pyloric sphincter, the intestine and the anus at 72 hpf).

This work resulted in producing multiple transcriptomic and chromatin accessibility datasets, development of the method to predict CRMs and TF binding sites within them, as well as the validation of the predictions through this method, which gave functional information on the transcription factors driving a certain predicted CRM (SpHox11/13b directly regulating *SpLox* through SpLoxCRM9). However, they are still incomplete, with the fullest dataset available only for *S. purpuratus,* generation of these datasets is in progress. Upon completion, all the aspects of a GRN controlling the gut patterning in the four deuterostomes can be deduced and compared between these species, allowing to highlight similarities and differences, contributing to the understanding of the evolution of this network. Therefore, this thesis, in addition to identification of the elements of a GRN upstream of *Lox* and *Cdx*, sets up a solid foundation for further research.

Non-book component

Non-book component contains the detailed results described in Chapters 3 and 4, and is presented on a USB stick along with a physical copy of this thesis. Structure of the USB stick is as follows:

```
C:
|   Information.txt
|
+---Branchiostoma lanceolatum
|   |   Bl_parahox_putative_CRMs.bed
|   |   Bl_putative_CRMs.bed
|   |
vertebrate_tf_motifs_Bl_parahox_putative_CRMs.bed
|   |
|   +---combined replicate narrowPeak files from Dr JL
Skarmeta Gomez and peak gene annotations
|   |       amphi_15h_peaks.bed
|   |       amphi_36h_peaks.bed
|   |       amphi_8h_peaks.bed
|   |       peakgenes_amphi_15h_peaks.txt
|   |       peakgenes_amphi_36h_peaks.txt
|   |       peakgenes_amphi_8h_peaks.txt
|   |
|   \---genome and annotations
|           Bl71nemr.fa
|           Bl_Annotation.gtf
|           gene_models_only_BraLan.gff3
|
+---Comparisons
|   +---motif PCA and Venn diagram input files
|   |       counts_cdx.txt
|   |       counts_lox.txt
|   |       counts_para.txt
|   |       venn_cdx.txt
|   |       venn_lox.txt
|   |       venn_para.txt
|   |
|   \---pair wise BLAST results
|       |   bl_parahox_crms.fa
|       |   pl_parahox_crms.fa
|       |   pm_parahox_crms.fa
|       |   sp_parahox_crms.fa
|       |
|       +---vsBl
|       |       vsBl_bl_crms.fa
|       |       vsBl_pl_crms.fa
|       |       vsBl_pm_crms.fa
|       |       vsBl_sp_crms.fa
|       |
|       +---vsPl
|       |       vsPl_bl_crms.fa
|       |       vsPl_pl_crms.fa
```

```
|       |       vsPl_pm_crms.fa
|       |       vsPl_sp_crms.fa
|       |
|       +---vsPm
|       |       vsPm_bl_crms.fa
|       |       vsPm_pl_crms.fa
|       |       vsPm_pm_crms.fa
|       |       vsPm_sp_crms.fa
|       |
|       \---vsSp
|               vsSp_bl_crms.fa
|               vsSp_pl_crms.fa
|               vsSp_pm_crms.fa
|               vsSp_sp_crms.fa
|
+---Paracentrotus lividus
|   |   Pl_parahox_putative_CRMs.bed
|   |   putative_CRMs_Pl.bed
|   |
vertebrate_tf_motifs_Pl_parahox_putative_CRMs.bed
|   |
|   +---combined replicates narrowPeak files and peak
gene annotations
|   |       peakgenes_pl_15hpf_wt.txt
|   |       peakgenes_pl_24hpf_wt.txt
|   |       peakgenes_pl_40hpf_wt.txt
|   |       pl_15hpf_wt.narrowPeak
|   |       pl_24hpf_wt.narrowPeak
|   |       pl_40hpf_wt.narrowPeak
|   |
|   +---genome and annotations
|   |       Pliv_PqN3S_evmp.gtf
|   |       Pliv_PqN3S_sm.fa
|   |
|   +---narrowPeak files
|   |       pl1_15hpf_wtA_peakcalling_peaks.narrowPeak
|   |       pl1_15hpf_wtB_peakcalling_peaks.narrowPeak
|   |       pl1_24hpf_wtA_peakcalling_peaks.narrowPeak
|   |       pl1_24hpf_wtB_peakcalling_peaks.narrowPeak
|   |       pl1_40hpf_wtA_peakcalling_peaks.narrowPeak
|   |       pl1_40hpf_wtB_peakcalling_peaks.narrowPeak
|   |
|   \---peakgenes
|           all_peak_genes.narrowPeak
|           for_graph_pl.tab
|           geneswpeaks.sh
|           gene_list.txt
|           peakgenes_pl_15hpf_wt.txt
|           peakgenes_pl_24hpf_wt.txt
|           peakgenes_pl_40hpf_wt.txt
|
+---Patiria miniata
|   |   Pm_parahox_putative_CRMs.bed
|   |   Pm_putative_CRMs.bed
|   |
vertebrate_tf_motifs_Pm_parahox_putative_CRMs.bed
|   |
|   +---combined replicates narrowPeak files and peak
gene annotations
```

```
|    |       peakgenes_pm_24hpf_wt.txt
|    |       peakgenes_pm_66hpf_wt.txt
|    |       peakgenes_pm_90hpf_wt.txt
|    |       pm_24hpf_wt.narrowPeak
|    |       pm_66hpf_wt.narrowPeak
|    |       pm_90hpf_wt.narrowPeak
|    |
|    +---genome and annotations
|    |       genes_Pm.gtf
|    |       pmin_scaffolds_v2.0.fa
|    |
|    \---narrowPeak files
|            pmin2_24hpf_wtA_peakcalling_peaks.narrowPeak
|            pmin2_24hpf_wtB_peakcalling_peaks.narrowPeak
|            pmin2_66hpf_wtA_peakcalling_peaks.narrowPeak
|            pmin2_66hpf_wtB_peakcalling_peaks.narrowPeak
|            pmin2_90hpf_wtA_peakcalling_peaks.narrowPeak
|            pmin2_90hpf_wtB_peakcalling_peaks.narrowPeak
|
\---Strongylocentrotus purpuratus
     |   Sp_parahox_putative_CRMs.bed
     |   Sp_putative_CRMs.bed
     |   Sp_tfs_in_parahox_CRMs_hindgut_72hpf.bed
     |
vertebrate_tf_motifs_Sp_parahox_putative_CRMs.bed
     |
     +---combined replicate narrowPeak files and peak
gene annotations
     |       peakgenes_sp_24hpf_wt.txt
     |       peakgenes_sp_48hpf_gut.txt
     |       peakgenes_sp_48hpf_wt.txt
     |       peakgenes_sp_66hpf_gut.txt
     |       peakgenes_sp_66hpf_wt.txt
     |       peakgenes_sp_72hpf_wt.txt
     |       sp_24hpf_wt.narrowPeak
     |       sp_48hpf_gut.narrowPeak
     |       sp_48hpf_wt.narrowPeak
     |       sp_66hpf_gut.narrowPeak
     |       sp_66hpf_wt.narrowPeak
     |       sp_72hpf_wt.narrowPeak
     |
     +---differential analysis
     |    |   named_48_RNA_UP_ATAC_UP.txt
     |    |   named_66_RNA_UP_ATAC_UP.txt
     |    |
     |    +---ATAC-seq
     |    |   +---48hpf
     |    |   |
denovo_homerResults_48hpf_atac_gut_more_open.html
     |    |   |       loci_sp48_atac_gut_DN.bed
     |    |   |       loci_sp48_atac_gut_UP.bed
     |    |   |       sp48_ATAC_gut_wt.tsv
     |    |   |
     |    |   \---66hpf
     |    |
denovo_homerResults_66hpf_gut_more_open.html
     |    |           loci_sp66_atac_gut_DN.bed
     |    |           loci_sp66_atac_gut_UP.bed
     |    |           sp66_ATAC_gut_wt.tsv
```

```
|   |
|   \---RNA-seq
|           annotated_Sp48hpf_gut_wt_RUV.csv
|           annotated_Sp66hpf_gut_wt_RUV.csv
|           annotated_Sp_gut66_gut48_RUV.csv
|
+---genome and annotations
|   |   SPU_function.txt
|   |   SPU_GO.txt
|   |   SPU_PFAM.txt
|   |   WHL_names_table.txt
|   |   WHL_NCBI.txt
|   |   WHL_SPU.txt
|   |
|   +---S. purpuratus genome 3.1
|   |       Spur_3.1.LinearScaffold.fa
|   |       Transcriptome.gtf
|   |
|   \---S.purpuratus genome 5.0
|           spur5.fasta
|           SPU_ids_on_spur5.gff3
|
+---HiC data
|   +---100kb resolution
|   |       chr11.mat
|   |       chr11_xgi.bed
|   |       sample1_100000.matrix
|   |       sample1_100000_abs.bed
|   |
|   \---20kb resolution ParaHox on target analysis
|           parahox_norm.mat
|           parahox_norm_xgi.bed
|           sample1_ontarget_20000.matrix
|           sample1_ontarget_20000_abs.bed
|           Sp_50hpf_pairs.txt
|           Sp_50hpf_significant_20000.txt
|
+---known CRMs
|       known_CRMs_TFs_check.txt
|       known_CRM_sea_urchin.xlsx
|
+---narrowPeak files
|       sp4_24hpf_wtA_peakcalling_peaks.narrowPeak
|       sp4_24hpf_wtB_peakcalling_peaks.narrowPeak
|       sp4_48hpf_gutA_peakcalling_peaks.narrowPeak
|       sp4_48hpf_gutB_peakcalling_peaks.narrowPeak
|       sp4_48hpf_wtA_peakcalling_peaks.narrowPeak
|       sp4_48hpf_wtB_peakcalling_peaks.narrowPeak
|       sp4_66hpf_gutA_peakcalling_peaks.narrowPeak
|       sp4_66hpf_gutB_peakcalling_peaks.narrowPeak
|       sp4_66hpf_wtA_peakcalling_peaks.narrowPeak
|       sp4_66hpf_wtB_peakcalling_peaks.narrowPeak
|       sp4_72hpf_wtA_peakcalling_peaks.narrowPeak
|       sp4_72hpf_wtB_peakcalling_peaks.narrowPeak
|
\---scRNAseq

        sp72_positive_genes_all_clusters.xlsx
        sp72_tf_info_all_clusters.xlsx
```

Publications

1. Magri, M.S., Voronov, D., Ranđelović, J., Cuomo, C., Gómez-Skarmeta, J.L. and Arnone, M.I., in press. ATAC-Seq for assaying chromatin accessibility protocol using echinoderm embryos. Methods in Molecular Biology

Contribution: adaptation of the protocol for the sea urchin embryos as well testing the protocol.

2. Lowe, E.K., Cuomo, C., Voronov, D. and Arnone, M.I., 2019. Using ATAC-seq and RNA-seq to increase resolution in GRN connectivity. Methods in Cell Biology, 151, p.115, https://doi.org/10.1016/bs.mcb.2018.11.001

This protocol is cited in multiple parts of this thesis, since it describes the major ideas and steps behind the GRN drafting using ATAC-seq and transcriptomic data. Thus, this thesis project has tested and tuned the described protocol.

Contribution: testing of the whole protocol, combination of the approach with different NGS data, write up of section 4 of the protocol.

Bibliography

Acemel, R.D. et al., 2016. A single three-dimensional chromatin compartment in amphioxus indicates a stepwise evolution of vertebrate Hox bimodal regulation. *Nature genetics*, 48(3), pp.336–341.

Adams, N.L. et al., 2019. Procuring animals and culturing of eggs and embryos. *Methods in cell biology*, 150, pp.3–46.

Akasaka, K. et al., 1994. Genomic organization of a gene encoding the spicule matrix protein SM30 in the sea urchin Strongylocentrotus purpuratus. *The Journal of biological chemistry*, 269(32), pp.20592–20598.

Altschul, S.F. et al., 1990. Basic local alignment search tool. *Journal of molecular biology*, 215(3), pp.403–410.

Andrews, S., 2010. *FastQC: a quality control tool for high throughput sequence data*, Available at: https://www.bioinformatics.babraham.ac.uk/projects/fastqc/.

Annunziata, R. et al., 2014. Pattern and process during sea urchin gut morphogenesis: the regulatory landscape. *Genesis* , 52(3), pp.251–268.

Annunziata, R. & Arnone, M.I., 2014. A dynamic regulatory network explains ParaHox gene control of gut patterning in the sea urchin. *Development* , 141(12), pp.2462–2472.

Annunziata, R., Martinez, P. & Arnone, M.I., 2013. Intact cluster and chordate-like expression of ParaHox genes in a sea star. *BMC biology*, 11, p.68.

Apweiler, R. et al., 2004. UniProt: the Universal Protein knowledgebase. *Nucleic acids research*, 32(Database issue), pp.D115–9.

Arenas-Mena, C., Cameron, R.A. & Davidson, E.H., 2006. Hindgut specification and cell-adhesion functions of Sphox11/13b in the endoderm of the sea urchin embryo. *Development, growth & differentiation*, 48(7), pp.463–472.

Arnone, M.I. et al., 2006. Genetic organization and embryonic expression of the ParaHox genes in the sea urchin S. purpuratus: insights into the relationship between clustering and colinearity. *Developmental biology*, 300(1), pp.63–73.

Arnone, M.I., Dmochowski, I.J. & Gache, C., 2004. Using reporter genes to study cis-regulatory elements. *Methods in cell biology*, 74, pp.621–652.

Arnone, M.I., Martin, E.L. & Davidson, E.H., 1998. Cis-regulation downstream of cell type specification: a single compact element controls the complex expression of the CyIIa gene in sea urchin embryos. *Development* , 125(8),

pp.1381–1395.

Barrington, E.J.W., 1937. VI - The digestive system of Amphioxus (Branchiostoma) Lanceolatus. *Philosophical transactions of the Royal Society of London. Series B, Biological sciences*, 228(553), pp.269–312.

Beckett, B.R. & Petkovich, M., 1999. Evolutionary Conservation in Retinoid Signalling and Metabolism. *American zoologist*, 39(4), pp.783–795.

Birnbaum, R.Y. et al., 2012. Coding exons function as tissue-specific enhancers of nearby genes. *Genome research*, 22(6), pp.1059–1068.

Blair, J.E. & Hedges, S.B., 2005. Molecular phylogeny and divergence times of deuterostome animals. *Molecular biology and evolution*, 22(11), pp.2275–2284.

Bolger, A.M., Lohse, M. & Usadel, B., 2014. Trimmomatic: a flexible trimmer for Illumina sequence data. *Bioinformatics* , 30(15), pp.2114–2120.

Brooke, N.M., Garcia-Fernàndez, J. & Holland, P.W., 1998. The ParaHox gene cluster is an evolutionary sister of the Hox gene cluster. *Nature*, 392(6679), pp.920–922.

Buenrostro, J.D. et al., 2015. ATAC-seq: A Method for Assaying Chromatin Accessibility Genome-Wide. *Current protocols in molecular biology / edited by Frederick M. Ausubel ... [et al.]*, 109, pp.21.29.1–9.

Buenrostro, J.D. et al., 2013. Transposition of native chromatin for fast and sensitive epigenomic profiling of open chromatin, DNA-binding proteins and nucleosome position. *Nature methods*, 10(12), pp.1213–1218.

Burke, R.D., 1981. Structure of the digestive tract of the pluteus larva of Dendraster excentricus (Echinodermata: Echinoida). *Zoomorphology*, 98(3), pp.209–225.

Calestani, C. & Rogers, D.J., 2010. Cis-regulatory analysis of the sea urchin pigment cell gene polyketide synthase. *Developmental biology*, 340(2), pp.249–255.

Cameron, R.A. et al., 2015. Do echinoderm genomes measure up? *Marine genomics*, 22, pp.1–9.

Cameron, R.A. et al., 2009. SpBase: the sea urchin genome database and web site. *Nucleic acids research*, 37(Database issue), pp.D750–4.

Carty, M. et al., 2017. An integrated model for detecting significant chromatin interactions from high-resolution Hi-C data. *Nature communications*, 8, p.15454.

Cary, G.A., Cameron, R.A. & Hinman, V.F., 2018. EchinoBase: Tools for Echinoderm Genome Analyses. *Methods in molecular biology* , 1757,

pp.349–369.

Cholley, P.-E. et al., 2018. Modeling gene-regulatory networks to describe cell fate transitions and predict master regulators. *NPJ systems biology and applications*, 4, p.29.

Coffman, J.A. et al., 1997. SpMyb functions as an intramodular repressor to regulate spatial expression of CyIIIa in sea urchin embryos. *Development* , 124(23), pp.4717–4727.

Cole, A.G. et al., 2009. Two ParaHox genes, SpLox and SpCdx, interact to partition the posterior endoderm in the formation of a functional gut. *Development* , 136(4), pp.541–549.

Cole, A.G. & Arnone, M.I., 2009. Fluorescent in situ hybridization reveals multiple expression domains for SpBrn1/2/4 and identifies a unique ectodermal cell type that co-expresses the ParaHox gene SpLox. *Gene expression patterns: GEP*, 9(5), pp.324–328.

Coulier, F. et al., 2000. Ancestrally-duplicated paraHOX gene clusters in humans. *International journal of oncology*, 17(3), pp.439–444.

Cui, M. et al., 2017a. Sequential Response to Multiple Developmental Network Circuits Encoded in an Intronic cis-Regulatory Module of Sea Urchin hox11/13b. *Cell reports*, 19(2), pp.364–374.

Cui, M. et al., 2017b. Sequential Response to Multiple Developmental Network Circuits Encoded in an Intronic cis- Regulatory Module of Sea Urchin hox11/13b. *Cell reports*, 19(2), pp.364–374.

Cui, Y. et al., 2016. BioCircos.js: an interactive Circos JavaScript library for biological data visualization on web applications. *Bioinformatics* , 32(11), pp.1740–1742.

Cusanovich, D.A. et al., 2018. A Single-Cell Atlas of In Vivo Mammalian Chromatin Accessibility. *Cell*, 174(5), pp.1309–1324.e18.

Damle, S. & Davidson, E.H., 2011. Precise cis-regulatory control of spatial and temporal expression of the alx-1 gene in the skeletogenic lineage of S. purpuratus. *Developmental biology*, 357(2), pp.505–517.

Davidson, E.H. et al., 2002. A provisional regulatory gene network for specification of endomesoderm in the sea urchin embryo. *Developmental biology*, 246(1), pp.162–190.

Davies, J.O.J. et al., 2017. How best to identify chromosomal interactions: a comparison of approaches. *Nature methods*, 14(2), pp.125–134.

Dehal, P. & Boore, J.L., 2005. Two rounds of whole genome duplication in the ancestral vertebrate. *PLoS biology*, 3(10), p.e314.

Dekker, J. et al., 2002. Capturing chromosome conformation. *Science*, 295(5558), pp.1306–1311.

Dekker, J., Marti-Renom, M.A. & Mirny, L.A., 2013. Exploring the three-dimensional organization of genomes: interpreting chromatin interaction data. *Nature reviews. Genetics*, 14(6), pp.390–403.

Donoghue, P.C.J. & Purnell, M.A., 2005. Genome duplication, extinction and vertebrate evolution. *Trends in ecology & evolution*, 20(6), pp.312–319.

Dudchenko, O. et al., 2017. De novo assembly of the genome using Hi-C yields chromosome-length scaffolds. *Science*, 356(6333), pp.92–95.

Duncan, S.A. et al., 1994. Expression of transcription factor HNF-4 in the extraembryonic endoderm, gut, and nephrogenic tissue of the developing mouse embryo: HNF-4 is a marker for primary endoderm in the implanting blastocyst. *Proceedings of the National Academy of Sciences of the United States of America*, 91(16), pp.7598–7602.

Dylus, D.V. et al., 2016. Large-scale gene expression study in the ophiuroid Amphiura filiformis provides insights into evolution of gene regulatory networks. *EvoDevo*, 7, p.2.

Eddy, S.R., 2011. Accelerated Profile HMM Searches. *PLoS computational biology*, 7(10), p.e1002195.

Ettensohn, C.A. et al., 2003. Alx1, a member of the Cart1/Alx3/Alx4 subfamily of Paired-class homeodomain proteins, is an essential component of the gene network controlling skeletogenic fate specification in the sea urchin embryo. *Development* , 130(13), pp.2917–2928.

Even-Faitelson, L. et al., 2016. Coming to terms with chromatin structure. *Chromosoma*, 125(1), pp.95–110.

Finn, R.D. et al., 2014. Pfam: the protein families database. *Nucleic acids research*, 42(Database issue), pp.D222–30.

Flores, R.L. & Livingston, B.T., 2017. The skeletal proteome of the sea star Patiria miniata and evolution of biomineralization in echinoderms. *BMC evolutionary biology*, 17(1), p.125.

Franks, R.R. et al., 1988. Direct introduction of cloned DNA into the sea urchin zygote nucleus, and fate of injected DNA. *Development* , 102(2), pp.287–299.

Fraser, J. et al., 2015. An Overview of Genome Organization and How We Got There: from FISH to Hi-C. *Microbiology and molecular biology reviews: MMBR*, 79(3), pp.347–372.

Garstang, M. & Ferrier, D.E.K., 2013. Time is of the essence for ParaHox homeobox gene clustering. *BMC biology*, 11, p.72.

Gaunt, S.J., 2015. The significance of Hox gene collinearity. *The International journal of developmental biology*, 59(4-6), pp.159–170.

Gaunt, S.J., Drage, D. & Cockley, A., 2003. Vertebrate caudal gene expression gradients investigated by use of chick cdx-A/lacZ and mouse cdx-1/lacZ reporters in transgenic mouse embryos: evidence for an intron enhancer. *Mechanisms of development*, 120(5), pp.573–586.

Gilbert, S.F., 2016. *Developmental Biology*, Sinauer.

Gildor, T. et al., 2019. Developmental transcriptomes of the sea star, Patiria miniata , illuminate the relationship between conservation of gene expression and morphological conservation. *Evolutionary Biology*.

Gildor, T. & Ben-Tabou de-Leon, S., 2015. Comparative Study of Regulatory Circuits in Two Sea Urchin Species Reveals Tight Control of Timing and High Conservation of Expression Dynamics. *PLoS genetics*, 11(7), p.e1005435.

Gildor, T., Hinman, V. & Ben-Tabou-De-Leon, S., 2017. Regulatory heterochronies and loose temporal scaling between sea star and sea urchin regulatory circuits. *The International journal of developmental biology*, 61(3-4-5), pp.347–356.

Graham, A., Papalopulu, N. & Krumlauf, R., 1989. The murine and Drosophila homeobox gene complexes have common features of organization and expression. *Cell*, 57(3), pp.367–378. Available at: http://dx.doi.org/10.1016/0092-8674(89)90912-4.

Heinz, S. et al., 2010. Simple Combinations of Lineage-Determining Transcription Factors Prime cis-Regulatory Elements Required for Macrophage and B Cell Identities. *Molecular Cell*, 38(4), pp.576–589. Available at: http://dx.doi.org/10.1016/j.molcel.2010.05.004.

Herzenberg, L.A., Sweet, R.G. & Herzenberg, L.A., 1976. Fluorescence-Activated Cell Sorting. *Scientific American*, 234(3), pp.108–117. Available at: http://dx.doi.org/10.1038/scientificamerican0376-108.

Holland, L.Z., 2015. Cephalochordata. In A. Wanninger, ed. *Evolutionary Developmental Biology of Invertebrates 6*. Advances in Experimental Medicine and Biology. Vienna: Springer Vienna, pp. 91–133.

Holland, L.Z. & Yu, J.-K., 2004. Cephalochordate (Amphioxus) Embryos: Procurement, Culture, and Basic Methods. In *Development of Sea Urchins, Ascidians, and Other Invertebrate Deuterostomes: Experimental Approaches*. Methods in Cell Biology. Elsevier, pp. 195–215.

Holland, P.W.H., 2013. Evolution of homeobox genes. *Wiley interdisciplinary reviews. Developmental biology*, 2(1), pp.31–45.

Holland, P.W.H., 2003. More genes in vertebrates? *Journal of structural and*

functional genomics, 3(1-4), pp.75–84.

Howard-Ashby, M. et al., 2006. Identification and characterization of homeobox transcription factor genes in Strongylocentrotus purpuratus, and their expression in embryonic development. *Developmental biology*, 300(1), pp.74–89.

Hwang, B., Lee, J.H. & Bang, D., 2018. Single-cell RNA sequencing technologies and bioinformatics pipelines. *Experimental & molecular medicine*, 50(8), p.96.

Ikuta, T. et al., 2013. Identification of an intact ParaHox cluster with temporal colinearity but altered spatial colinearity in the hemichordate Ptychodera flava. *BMC evolutionary biology*, 13, p.129.

Jakubison, B.L. et al., 2018. Induced PTF1a expression in pancreatic ductal adenocarcinoma cells activates acinar gene networks, reduces tumorigenic properties, and sensitizes cells to gemcitabine treatment. *Molecular oncology*, 12(7), pp.1104–1124.

Ji, C. et al., 2012. Echinoderms have bilateral tendencies. *PloS one*, 7(1), p.e28978.

John, S. et al., 2011. Chromatin accessibility pre-determines glucocorticoid receptor binding patterns. *Nature genetics*, 43(3), pp.264–268.

Juliano, C., Swartz, S.Z. & Wessel, G., 2014. Isolating specific embryonic cells of the sea urchin by FACS. *Methods in molecular biology* , 1128, pp.187–196.

Kam, R.K.T. et al., 2012. Retinoic acid synthesis and functions in early embryonic development. *Cell & bioscience*, 2(1), p.11.

Khan, A. et al., 2018. JASPAR 2018: update of the open-access database of transcription factor binding profiles and its web framework. *Nucleic acids research*, 46(D1), pp.D260–D266.

Kim, J. et al., 2019. Clinical Factors Associated With Gastric Cancer in Individuals with Lynch Syndrome. *Clinical gastroenterology and hepatology: the official clinical practice journal of the American Gastroenterological Association*. Available at: http://dx.doi.org/10.1016/j.cgh.2019.07.012.

Kishimura, H. & Hayashi, K., 2005. Characterization of phospholipase A2 from the pyloric ceca of two species of starfish, Coscinasterias acutispina and Plazaster borealis. *Food chemistry*, 92(3), pp.407–411.

Kostrewa, D. et al., 2009. RNA polymerase II-TFIIB structure and mechanism of transcription initiation. *Nature*, 462(7271), pp.323–330.

Kudtarkar, P. & Cameron, R.A., 2017. Echinobase: an expanding resource for echinoderm genomic information. *Database: the journal of biological*

databases and curation, 2017. Available at: http://dx.doi.org/10.1093/database/bax074.

Langmead, B. & Salzberg, S.L., 2012. Fast gapped-read alignment with Bowtie 2. *Nature methods*, 9(4), pp.357–359.

Lee, P.Y., Nam, J. & Davidson, E.H., 2007. Exclusive developmental functions of gatae cis-regulatory modules in the Strongylocentrorus purpuratus embryo. *Developmental biology*, 307(2), pp.434–445.

de-Leon, S.B.-T. & Davidson, E.H., 2010. Information processing at the foxa node of the sea urchin endomesoderm specification network. *Proceedings of the National Academy of Sciences of the United States of America*, 107(22), pp.10103–10108.

Levine, M. & Tjian, R., 2003. Transcription regulation and animal diversity. *Nature*, 424(6945), pp.147–151. Available at: http://dx.doi.org/10.1038/nature01763.

Lieberman-Aiden, E. et al., 2009. Comprehensive mapping of long-range interactions reveals folding principles of the human genome. *Science*, 326(5950), pp.289–293.

Li, M. et al., 2015. Medaka vasa gene has an exonic enhancer for germline expression. *Gene*, 555(2), pp.403–408.

Ling, M.M. & Robinson, B.H., 1997. Approaches to DNA mutagenesis: an overview. *Analytical biochemistry*, 254(2), pp.157–178.

Livi, C.B. & Davidson, E.H., 2006. Expression and function of blimp1/krox, an alternatively transcribed regulatory gene of the sea urchin endomesoderm network. *Developmental Biology*, 293(2), pp.513–525. Available at: http://dx.doi.org/10.1016/j.ydbio.2006.02.021.

Longabaugh, W.J.R., Davidson, E.H. & Bolouri, H., 2005. Computational representation of developmental genetic regulatory networks. *Developmental biology*, 283(1), pp.1–16.

Love, M.I., Huber, W. & Anders, S., 2014. Moderated estimation of fold change and dispersion for RNA-seq data with DESeq2. *Genome biology*, 15(12), p.550.

Lowe, C.J. et al., 2015. The deuterostome context of chordate origins. *Nature*, 520(7548), pp.456–465.

Lowe, E.K. et al., 2019. Using ATAC-seq and RNA-seq to increase resolution in GRN connectivity. *Methods in cell biology*, 151, pp.115–126.

Lowe, E.K., Cuomo, C. & Arnone, M.I., 2016. A Differential Transcriptomic Approach to Compare Target Genes of Homologous Transcription Factors in Echinoderm Species. *Dynamics of Mathematical Models in Biology*,

pp.55–63. Available at: http://dx.doi.org/10.1007/978-3-319-45723-9_5.

Lowe, E.K., Cuomo, C. & Arnone, M.I., 2017. Omics approaches to study gene regulatory networks for development in echinoderms. *Briefings in Functional Genomics*, 16(5), pp.299–308. Available at: http://dx.doi.org/10.1093/bfgp/elx012.

Luo, Y.-J. & Su, Y.-H., 2012. Opposing nodal and BMP signals regulate left-right asymmetry in the sea urchin larva. *PLoS biology*, 10(10), p.e1001402.

Magri, M.S. et al., ATAC-Seq for assaying chromatin accessibility protocol using echinoderm embryos. In D. Carroll & S. Stricker, eds. *Developmental Biology of the Sea Urchin and Other Marine Invertebrates*. Methods in Molecular Biology. Humana Press.

Makabe, K.W. et al., 1995. Cis-regulatory control of the SM50 gene, an early marker of skeletogenic lineage specification in the sea urchin embryo. *Development* , 121(7), pp.1957–1970.

Malik, A. et al., 2017. Parallel embryonic transcriptional programs evolve under distinct constraints and may enable morphological conservation amidst adaptation. *Developmental biology*, 430(1), pp.202–213.

Marlétaz, F. et al., 2018. Amphioxus functional genomics and the origins of vertebrate gene regulation. *Nature*, 564(7734), pp.64–70.

Materna, S.C., 2017. Using Morpholinos to Probe Gene Networks in Sea Urchin. *Methods in molecular biology* , 1565, pp.87–104.

Materna, S.C., Swartz, S.Z. & Smith, J., 2013. Notch and Nodal control forkhead factor expression in the specification of multipotent progenitors in sea urchin. *Development* , 140(8), pp.1796–1806.

McCarty, C.M. & Coffman, J.A., 2013. Developmental cis-regulatory analysis of the cyclin D gene in the sea urchin Strongylocentrotus purpuratus. *Biochemical and biophysical research communications*, 440(3), pp.413–418.

McClay, D.R., 2004. Methods for embryo dissociation and analysis of cell adhesion. *Methods in cell biology*, 74, pp.311–329.

McGinnis, W. & Krumlauf, R., 1992. Homeobox genes and axial patterning. *Cell*, 68(2), pp.283–302.

Metzis, V. et al., 2018. Nervous System Regionalization Entails Axial Allocation before Neural Differentiation. *Cell*, 175(4), pp.1105–1118.e17.

Mifsud, B. et al., 2017. GOTHiC, a probabilistic model to resolve complex biases and to identify real interactions in Hi-C data. *PloS one*, 12(4), p.e0174744.

Minokawa, T., Wikramanayake, A.H. & Davidson, E.H., 2005. cis-Regulatory inputs of the wnt8 gene in the sea urchin endomesoderm network. *Developmental biology*, 288(2), pp.545–558.

Moreno, E. et al., 2009. Tracking the origins of the bilaterian Hox patterning system: insights from the acoel flatworm Symsagittifera roscoffensis. *Evolution & development*, 11(5), pp.574–581.

Moses, A.M. et al., 2006. Large-scale turnover of functional transcription factor binding sites in Drosophila. *PLoS computational biology*, 2(10), p.e130.

Nakayama, S., Sekiguchi, T. & Ogasawara, M., 2019. Molecular and evolutionary aspects of the protochordate digestive system. *Cell and tissue research*. Available at: http://dx.doi.org/10.1007/s00441-019-03035-5.

Nam, J. et al., 2007. Cis-regulatory control of the nodal gene, initiator of the sea urchin oral ectoderm gene network. *Developmental biology*, 306(2), pp.860–869.

Nam, J. et al., 2010. Functional cis-regulatory genomics for systems biology. *Proceedings of the National Academy of Sciences*, 107(8), pp.3930–3935. Available at: http://dx.doi.org/10.1073/pnas.1000147107.

Neznanov, N., Umezawa, A. & Oshima, R.G., 1997. A regulatory element within a coding exon modulates keratin 18 gene expression in transgenic mice. *The Journal of biological chemistry*, 272(44), pp.27549–27557.

Oliveri, P., Tu, Q. & Davidson, E.H., 2008. Global regulatory logic for specification of an embryonic cell lineage. *Proceedings of the National Academy of Sciences of the United States of America*, 105(16), pp.5955–5962.

Ong, C.-T. & Corces, V.G., 2011. Enhancer function: new insights into the regulation of tissue-specific gene expression. *Nature reviews. Genetics*, 12(4), pp.283–293.

Osborne, C.S. et al., 2004. Active genes dynamically colocalize to shared sites of ongoing transcription. *Nature genetics*, 36(10), pp.1065–1071.

Osborne, P.W. et al., 2009. Differential regulation of ParaHox genes by retinoic acid in the invertebrate chordate amphioxus (Branchiostoma floridae). *Developmental biology*, 327(1), pp.252–262.

Oulhen, N. & Wessel, G.M., 2016. Albinism as a visual, in vivo guide for CRISPR/Cas9 functionality in the sea urchin embryo. *Molecular reproduction and development*, 83(12), pp.1046–1047.

Patro, R. et al., 2017. Salmon provides fast and bias-aware quantification of transcript expression. *Nature methods*, 14(4), pp.417–419.

Peichel, C.L. et al., 2017. Improvement of the Threespine Stickleback Genome

Using a Hi-C-Based Proximity-Guided Assembly. *The Journal of heredity*, 108(6), pp.693–700.

Perez-Villamil, B., Schwartz, P.T. & Vallejo, M., 1999. The pancreatic homeodomain transcription factor IDX1/IPF1 is expressed in neural cells during brain development. *Endocrinology*, 140(8), pp.3857–3860.

Perillo, M. et al., 2016. A pancreatic exocrine-like cell regulatory circuit operating in the upper stomach of the sea urchin Strongylocentrotus purpuratus larva. *BMC evolutionary biology*, 16(1), p.117.

Perillo, M. et al., 2018. New Neuronal Subtypes With a “Pre-Pancreatic” Signature in the Sea Urchin. *Frontiers in endocrinology*, 9, p.650.

Peter, I.S. & Davidson, E.H., 2015. *Genomic Control Process: Development and Evolution*, Elsevier Science.

Peter, I.S. & Davidson, E.H., 2010. The endoderm gene regulatory network in sea urchin embryos up to mid-blastula stage. *Developmental biology*, 340(2), pp.188–199.

Putnam, N.H. et al., 2008. The amphioxus genome and the evolution of the chordate karyotype. *Nature*, 453(7198), pp.1064–1071.

Quinlan, A.R. & Hall, I.M., 2010. BEDTools: a flexible suite of utilities for comparing genomic features. *Bioinformatics* , 26(6), pp.841–842.

Ransick, A. & Davidson, E.H., 2006. cis-regulatory processing of Notch signaling input to the sea urchin glial cells missing gene during mesoderm specification. *Developmental biology*, 297(2), pp.587–602.

Revilla-i-Domingo, R., Minokawa, T. & Davidson, E.H., 2004. R11: a cis-regulatory node of the sea urchin embryo gene network that controls early expression of SpDelta in micromeres. *Developmental biology*, 274(2), pp.438–451.

Rhoads, A. & Au, K.F., 2015. PacBio Sequencing and Its Applications. *Genomics, proteomics & bioinformatics*, 13(5), pp.278–289.

Risso, D. et al., 2014. Normalization of RNA-seq data using factor analysis of control genes or samples. *Nature Biotechnology*, 32(9), pp.896–902. Available at: http://dx.doi.org/10.1038/nbt.2931.

Ritchie, M.E. et al., 2015. limma powers differential expression analyses for RNA-sequencing and microarray studies. *Nucleic acids research*, 43(7), p.e47.

Ritter, D.I. et al., 2012. Transcriptional enhancers in protein-coding exons of vertebrate developmental genes. *PloS one*, 7(5). Available at: http://dx.doi.org/10.1371/journal.pone.0035202.

Robinson, J.T. et al., 2011. Integrative genomics viewer. *Nature biotechnology*, 29(1), pp.24–26.

Roy, S. & Kundu, T.K., 2014. Gene regulatory networks and epigenetic modifications in cell differentiation. *IUBMB life*, 66(2), pp.100–109.

Sea Urchin Genome Sequencing Consortium et al., 2006. The genome of the sea urchin Strongylocentrotus purpuratus. *Science*, 314(5801), pp.941–952.

Servant, N. et al., 2015. HiC-Pro: an optimized and flexible pipeline for Hi-C data processing. *Genome biology*, 16, p.259.

Servant, N. et al., 2012. HiTC: exploration of high-throughput “C” experiments. *Bioinformatics* , 28(21), pp.2843–2844.

Shashikant, T. & Ettensohn, C.A., 2019. Genome-wide analysis of chromatin accessibility using ATAC-seq. *Methods in cell biology*, 151, pp.219–235.

Shashikant, T., Khor, J.M. & Ettensohn, C.A., 2018. Global analysis of primary mesenchyme cell cis-regulatory modules by chromatin accessibility profiling. , pp.1–18.

Shlyueva, D., Stampfel, G. & Stark, A., 2014. Transcriptional enhancers: from properties to genome-wide predictions. *Nature reviews. Genetics*, 15(4), pp.272–286.

Siebert, S. et al., 2018. Stem cell differentiation trajectories in Hydra resolved at single-cell resolution. *bioRxiv*. Available at: http://dx.doi.org/10.1101/460154.

Singh, H., Khan, A.A. & Dinner, A.R., 2014. Gene regulatory networks in the immune system. *Trends in immunology*, 35(5), pp.211–218.

Smith, J. et al., 2008. A spatially dynamic cohort of regulatory genes in the endomesodermal gene network of the sea urchin embryo. *Developmental biology*, 313(2), pp.863–875.

Solek, C.M. et al., 2013. An ancient role for Gata-1/2/3 and Scl transcription factor homologs in the development of immunocytes. *Developmental biology*, 382(1), pp.280–292.

Strathmann, R.R., 1993. Hypotheses on the Origins of Marine Larvae. *Annual review of ecology and systematics*, 24(1), pp.89–117.

Stuart, T. et al., 2019. Comprehensive Integration of Single-Cell Data. *Cell*, 177(7), pp.1888–1902.e21.

Talbert, P.B., Meers, M.P. & Henikoff, S., 2019. Old cogs, new tricks: the evolution of gene expression in a chromatin context. *Nature reviews. Genetics*, 20(5), pp.283–297.

Trapnell, C. et al., 2010. Transcript assembly and quantification by RNA-Seq reveals unannotated transcripts and isoform switching during cell differentiation. *Nature biotechnology*, 28(5), pp.511–515.

Tsompana, M. & Buck, M.J., 2014. Chromatin accessibility: a window into the genome. *Epigenetics & chromatin*, 7(1), p.33.

Tu, Q. et al., 2012. Gene structure in the sea urchin Strongylocentrotus purpuratus based on transcriptome analysis. *Genome research*, 22(10), pp.2079–2087.

Tu, Q. et al., 2006. Sea urchin Forkhead gene family: phylogeny and embryonic expression. *Developmental biology*, 300(1), pp.49–62.

Tu, Q., Andrew Cameron, R. & Davidson, E.H., 2014. Quantitative developmental transcriptomes of the sea urchin Strongylocentrotus purpuratus. *Developmental Biology*, 385(2), pp.160–167. Available at: http://dx.doi.org/10.1016/j.ydbio.2013.11.019.

Ulianov, S.V., Gavrilov, A.A. & Razin, S.V., 2015. Nuclear compartments, genome folding, and enhancer-promoter communication. *International review of cell and molecular biology*, 315, pp.183–244.

Untergasser, A. et al., 2012. Primer3--new capabilities and interfaces. *Nucleic acids research*, 40(15), p.e115.

Vymetalkova, V. et al., 2019. DNA methylation and chromatin modifiers in colorectal cancer. *Molecular aspects of medicine*. Available at: http://dx.doi.org/10.1016/j.mam.2019.04.002.

Wahl, M.E. et al., 2009. The cis-regulatory system of the tbrain gene: Alternative use of multiple modules to promote skeletogenic expression in the sea urchin embryo. *Developmental biology*, 335(2), pp.428–441.

Wang, Y.-M. et al., 2012. Correlation between DNase I hypersensitive site distribution and gene expression in HeLa S3 cells. *PloS one*, 7(8), p.e42414.

Wickham, H., 2016. *ggplot2: Elegant Graphics for Data Analysis*, Springer.

Wikramanayake, A.H. et al., 2004. Nuclear beta-catenin-dependent Wnt8 signaling in vegetal cells of the early sea urchin embryo regulates gastrulation and differentiation of endoderm and mesodermal cell lineages. *Genesis* , 39(3), pp.194–205.

Wu, T.D. & Watanabe, C.K., 2005. GMAP: a genomic mapping and alignment program for mRNA and EST sequences. *Bioinformatics* , 21(9), pp.1859–1875.

Xiong, A.-S. et al., 2006. PCR-based accurate synthesis of long DNA sequences. *Nature protocols*, 1(2), pp.791–797.

Yaguchi, J., 2019. Microinjection methods for sea urchin eggs and blastomeres. *Methods in cell biology*, 150, pp.173–188.

Yaguchi, J. & Yaguchi, S., 2019. Evolution of nitric oxide regulation of gut function. *Proceedings of the National Academy of Sciences*, 2019(13), p.201816973.

Yang, X., Feng, S. & Tang, K., 2017. COUP - TF Genes, Human Diseases, and the Development of the Central Nervous System in Murine Models. In *Nuclear Receptors in Development and Disease*. Current Topics in Developmental Biology. Elsevier, pp. 275–301.

Yuh, C.-H. et al., 2004. An otx cis -regulatory module: a key node in the sea urchin endomesoderm gene regulatory network. *Developmental biology*, 269(2), pp.536–551.

Yuh, C.-H. et al., 2002. Patchy Interspecific Sequence Similarities Efficiently Identify Positive cis-Regulatory Elements in the Sea Urchin. *Developmental biology*, 246(1), pp.148–161.

Yuh, C.H., Bolouri, H. & Davidson, E.H., 2001. Cis-regulatory logic in the endo16 gene: switching from a specification to a differentiation mode of control. *Development* , 128(5), pp.617–629.

Yuh, C.-H., Dorman, E.R. & Davidson, E.H., 2005. Brn1/2/4, the predicted midgut regulator of the endo16 gene of the sea urchin embryo. *Developmental biology*, 281(2), pp.286–298.

Yu, J.K.S. & Holland, L.Z., 2009. Cephalochordates (Amphioxus or Lancelets): A Model for Understanding the Evolution of Chordate Characters. *Cold Spring Harbor protocols*, 2009(9), p.db.emo130–pdb.emo130.

Zhang, Y. et al., 2008. Model-based analysis of ChIP-Seq (MACS). *Genome biology*, 9(9), p.R137.

Zhao, Z. et al., 2006. Circular chromosome conformation capture (4C) uncovers extensive networks of epigenetically regulated intra- and interchromosomal interactions. *Nature genetics*, 38(11), pp.1341–1347.